

Stochastical Models for Networks in the Life Sciences

DISSERTATION

zur Erlangung des akademischen Grades
doctor rerum naturalium
(Dr. rer. nat.)
im Fach Informatik

eingereicht an der
Mathematisch-Naturwissenschaftlichen Fakultät II
der Humboldt-Universität zu Berlin

von
Herrn Dipl.-Inf. Michael Behrisch
geboren am 13.07.1976 in Berlin

Präsident der Humboldt-Universität zu Berlin:
Prof. Dr. Christoph Marksches

Dekan der Mathematisch-Naturwissenschaftlichen Fakultät II:
Prof. Dr. Wolfgang Coy

Gutachter:

1. Prof. Dr. Hans Jürgen Prömel
2. Prof. Dr. Anuschirawan Taraz
3. Priv.-Doz. Dr. Amin Coja-Oghlan

eingereicht am:	1. Dezember 2006
Tag der mündlichen Prüfung:	23. April 2007

Abstract

Motivated by structural properties of molecular similarity networks we study the behaviour of the component evolution in two different stochastic network models, that is random hypergraphs and random intersection graphs.

We prove gaussian distribution for the number of vertices in the giant component of a random d -uniform hypergraph (a local limit theorem in the $H_d(n, p)$ model for $p = c/\binom{n-1}{d-1}$ with $(d-1)^{-1} + \varepsilon < c < \infty$). We provide a proof using only probabilistic arguments, avoiding enumerative methods completely. This fundamental result is followed by further limit theorems concerning joint distributions of vertices and edges as well as the connectivity probability of random hypergraphs and the number of connected hypergraphs.

Due to deficiencies of the hypergraph model in reflecting properties of the real-world data, we switch the model and study the evolution of the order of the largest component in the random intersection graph model which reflects some clustering properties of real-world networks. We show that for appropriate choice of the parameters random intersection graphs differ from random (hyper-)graphs in that neither the so-called giant component, appearing when the average number of neighbours of a vertex gets larger than one, has linear order nor is the second largest of logarithmic order in the number of vertices.

Furthermore we describe a polynomial time algorithm for covering graphs with cliques, prove its asymptotic optimality in a random intersection graph model and study the evolution of the chromatic number in the model showing that, in a certain range of parameters, these random graphs can be coloured optimally with high probability using different greedy algorithms. Experiments on real network data confirm the positive theoretical predictions and suggest that heuristics for the clique and the chromatic number can work hand in hand proving mutual optimality.

Keywords:

random graph, giant component, intersection graph, complex network

Zusammenfassung

Motiviert durch strukturelle Eigenschaften molekularer Ähnlichkeitsnetzwerke werden die Evolution der größten Komponente eines Netzwerkes in zwei verschiedenen stochastischen Modellen, zufälligen Hypergraphen und zufälligen Schnittgraphen, untersucht.

Zuerst wird bewiesen, dass die Anzahl der Knoten in der größten Komponente d -uniformer Hypergraphen einer Normalverteilung folgt (lokaler Grenzwertsatz für das binomiale Zufallsmodell $H_d(n, p)$ für $p = c/\binom{n-1}{d-1}$ mit $(d-1)^{-1} + \varepsilon < c < \infty$). Der Beweis nutzt dabei ausschließlich probabilistische Argumente und keine enumerative Kombinatorik. Diesem grundlegenden Resultat folgen weitere Grenzwertsätze für die gemeinsame Verteilung von Knoten- und Kantenzahl sowie Sätze zur Zusammenhangswahrscheinlichkeit zufälliger Hypergraphen und zur asymptotischen Anzahl zusammenhängender Hypergraphen.

Da das Hypergraphenmodell einige Eigenschaften der Realweltdaten nur unzureichend abbildet, wird anschließend die Evolution der größten Komponente in zufälligen Schnittgraphen, die einige Clustereigenschaften realer Netzwerke gut widerspiegeln, untersucht. Es wird gezeigt, dass bei geeigneter Wahl der Parameter zufällige Schnittgraphen sich von zufälligen (Hyper-)Graphen dadurch unterscheiden, dass bei Erreichen einer durchschnittlichen Anzahl von Nachbarn von mehr als eins weder eine größte Komponente linearer Größe existiert, noch die zweitgrößte Komponente von logarithmischer Größe in Abhängigkeit von der Knotenzahl ist.

Weiterhin wird ein Polynomialzeitalgorithmus zur Überdeckung der Kanten eines Graphen mit möglichst wenigen Cliques (vollständigen Graphen) beschrieben und seine asymptotische Optimalität im Modell der zufälligen Schnittgraphen bewiesen. Anschließend wird die Entwicklung der chromatischen Zahl untersucht und gezeigt, dass, bei geeigneter Wahl der Parameter, zufällige Schnittgraphen mit hoher Wahrscheinlichkeit mittels verschiedener Greedystrategien optimal gefärbt werden können. Letztendlich zeigen Experimente auf realen Netzen eine Übereinstimmung mit den theoretischen Vorhersagen und legen eine gegenseitige Zertifizierung der Optimalität von Cliques- und Färbungszahl durch Heuristiken nahe.

Schlagwörter:

zufälliger Graph, große Komponente, Schnittgraph, komplexes Netzwerk

To Birgit, Miria and Amos

Contents

Preface	xi
I Random Hypergraphs and their Giant Component	1
1 Introduction	3
1.1 Related work	5
1.2 Techniques and outline.	6
1.3 Preliminaries	9
1.3.1 The Phase Transition and the Giant Component	11
2 A Central Limit Theorem for the Number of Vertices	13
2.1 Results	13
2.2 Stein's Method for Random Hypergraphs	14
2.3 Conditions for the Normality of $\mathcal{N}(H_d(n, p))$	18
2.4 An Upper Bound for δ	21
3 A Local Limit Theorem for the Number of Vertices	27
3.1 Results	27
3.2 Proof of the Local Limit Theorem	28
3.2.1 Outline	28
3.2.2 The Distribution of \mathcal{N}_3 as a Combination of \mathcal{N}_1 and \mathcal{S}	30
3.3 The Conditional Distribution of \mathcal{S}	31
3.3.1 Outline	31
3.3.2 Locality of \mathcal{S}_G	34
3.3.3 Approximating \mathcal{S} via \mathcal{S}_G	36
3.3.4 The Expectation of \mathcal{S}_G	38
3.3.5 The Variance of \mathcal{S}_G	40
3.3.6 The Number of Attached Isolated Vertices	44
3.4 Central Limit Theorem for \mathcal{S}	46
4 Bivariate Limit Theorems	49
4.1 Results	49
4.2 Bivariate Limit Theorem	51

4.2.1	Outline	51
4.2.2	Fourier Analysis	55
4.2.3	An Explicit Formula for the $H_d(n, p_z)$ Distribution $f(z)$	57
4.2.4	Continuity of $g(z)$	59
4.2.5	Proof of Lemma 4.16	61
4.2.6	Convolution	63
4.3	Calculations	64
4.3.1	The Distribution for $H_d(n, m)$	65
4.3.2	The Distribution for $H_d(n, p)$	74
5	Applications of the Local Limit Theorems	81
5.1	Results	81
5.1.1	The Probability of Connectedness	81
5.1.2	The Distribution of $\mathcal{M}(H_d(n, p))$ given Connectedness	83
5.2	Techniques	83
5.3	Probability of Connectedness in the Binomial Model	84
5.4	Connectivity Probability and the Number of Connected Graphs	90
5.5	Edge Distribution of Connected Hypergraphs	94
II	Random Intersection Graphs	97
6	Introduction	99
6.1	A Different Model for Random Graphs	100
6.1.1	Intersection Graphs	100
6.1.2	Random Intersection Graphs	100
6.1.3	Related Work	101
6.1.4	Overview	101
6.2	Auxiliary lemmas	102
7	Component Evolution	105
7.1	Results	105
7.2	Branching Processes	106
7.3	The Evolution for $\alpha > 1$	107
7.3.1	The Size of the Feature Set	107
7.3.2	Proof of Theorem 7.1, (7.1) and (7.2)	108
7.4	The Evolution for $\alpha < 1$	111
7.4.1	Feature Cliques as Components	113
8	Clique cover and feature reconstruction	115
8.1	Results	115
8.2	The Algorithm	116
8.3	The case $k = 1$	118
8.4	The case $k > 1$	119

9	Colouring heuristics and the clique number	123
9.1	Results	123
9.2	Proofs	125
9.2.1	Perfect Elimination Scheme	125
9.2.2	Smallest Last Heuristic	127
10	Experiments	133
10.1	The Giant Component	133
10.2	The Networks	135
10.3	Clique Cover	136
10.4	Colouring	139
11	Conclusion and Outlook	143
11.1	Random Intersection Graphs	143
11.2	Clique Cover	144
11.3	Colouring and Independence Number	144
11.4	Diameter	145
11.5	Clustering Coefficient	145
11.6	Final Remarks	146

Preface

Imagine a huge database of molecules which may be for instance drug substances or parts of proteins. One of the essential challenges in nowadays attempts of in-silico drug design and protein structure revelation is not only to collect these data but to arrange it in a form which makes it accessible for further manipulation, searches etc. One simple example is the search for relatives (i. e. similar structures) in such a huge database, for instance in order to find substances which avoid certain side effects while having the same effects.

To perform such a search efficiently it is very useful to have knowledge about the inner structure of the relationship network. Are there many small islands of strongly similar molecules which are more or less isolated from one another or is similarity a result of pure randomness i.e. the similarity links are scattered over the whole database?

The starting point of this thesis was a striking effect which was observed in the evolution of such a similarity network. Slowly lowering the threshold for what we call similar thereby inserting more and more similarity links (edges) between the molecules (vertices) we studied the connectivity structure, especially the number of vertices in the largest component (where a component is a set of vertices which are mutually reachable by following the links of the network). What we found was that there are essentially three episodes in this evolution, one where the largest component is quite small, one where it grows slowly and the third (after a sudden jump) where it covers almost the whole graph. Figure 1 shows the inverse evolution, that is, at the beginning all edges are inside the graph and going along the x-axis the threshold for similarity raises. Thus edges are removed and the components become smaller.

While the fact that a jump occurs was already known, reflecting the so-called threshold behaviour of a number of properties in random graphs, our aim was now to find explanations for the slow growth in the beginning of the evolution thereby gaining deeper understanding of the nature of the networks which will enable us to design algorithms specific to the networks and even prove their optimality.

The construction of stochastic models for complex real-world networks of huge dimensions has attracted an enormous amount of attention during the last five years. These efforts are motivated by several aspects, namely the prediction of network structure as well as the design, benchmarking and theoretical verification of algorithms.

As *graphs* are the canonical model for networks, *random graphs* seem to be appropriate candidates for the stochastic models. The first object to probabilistic studies was the

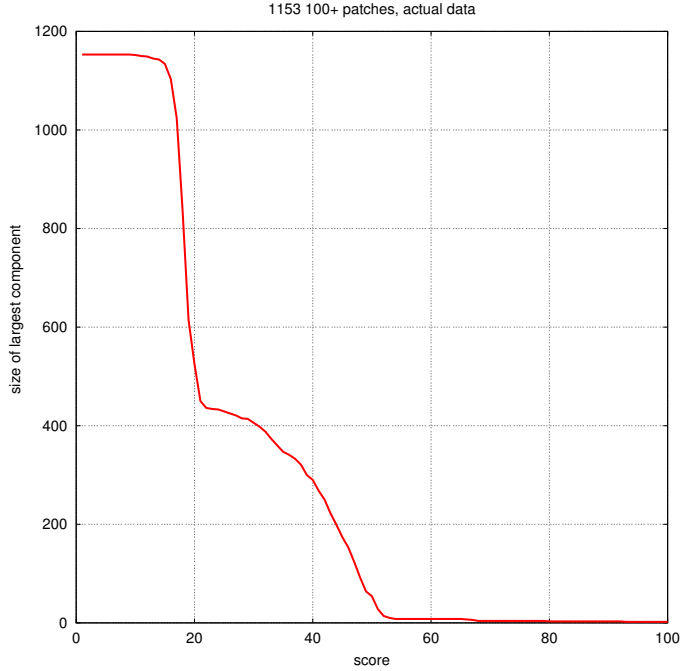


Figure 1: Largest component in the protein graph

classical random graph model introduced by Erdős and Rényi in the late 1950s. It is denoted by $G_{n,p}$ and considers a fixed set of n vertices and edges that exist with a certain probability $p = p(n)$, independently from each other. There exist variants of their model which fix the number of edges in advance ($G_{n,m}$ denotes the graph chosen uniformly at random among all graphs having n vertices and m edges) or allow hyperedges (edges containing more than two vertices).

Looking at our starting point, the evolution of the largest component, Erdős-Rényi-Graphs seem to be very well studied, thus we focused on d -uniform hypergraphs (all edges have d vertices) which include the standard graphs as a special case and on the range of p where the so-called giant component appears. There we can give general results on the asymptotic distribution of the number of vertices and the number of edges in the giant component in the binomial ($H_d(n, p)$) as well as in the uniform ($H_d(n, m)$) model. The precision achieved in the estimations (we get a Local Limit Theorem for the joint distribution) helps in solving further problems such as calculating the asymptotic number of connected graphs with a given number of vertices and edges.

Unfortunately it turns out that those graphs are not very well suited to model the behaviour of the largest component in real-world networks such as protein interaction networks or the WWW. Thus we turn to another random graph model which tries to reflect the special properties of our real-world instances. One of those properties is the transitivity which is inherent to similarity networks, since the similarity of molecule A and molecule B together with the similarity of B and C should obviously at least increase

the probability of a similarity between A and C. This is the main motivation for the study of random intersection graphs in Part II, where vertices get connected according to common features (assigned to them at random) which reflects the transitivity issues mentioned.

After the study of the evolution of the giant component in the new model (Chapter 7), which reflects the special behaviour (at least qualitatively) of the real-world network considerably better we turn to the study of optimisation problems on those graphs e.g. clique cover (Chapter 8) and colouring (Chapter 9). The reason for studying those two problems is mainly that a clique cover of an intersection graph can give insight into the semantics of the links in the net by assuming that a single feature is responsible for one clique in the cover. The colouring problem is one of the most studied optimisation problems in graph theory and can give a first insight on the difficulty of optimisation problems on intersection graphs while at the same time it certifies results from the clique cover problem as being optimal. We close the second part with experimental results and an outlook of open problems on random intersection graphs.

Part I of this thesis presents most of the results obtained in Behrisch, Coja-Oghlan and Kang [2006a,b], while the results of Part II are covered by Behrisch [2006], Behrisch and Taraz [2006], Behrisch et al. [2005].

While both parts are essential steps in the search for better stochastic models for real-world networks we tried to keep them as self-contained as possible because the readership attracted might be different for both parts. Thus while we tried to avoid overloading of symbols and diverging definitions of terms some basic concepts and utilities will be defined twice.

Every part starts with an introductory chapter fixing notation, giving an account on the related work and stating some auxiliary results while the following chapters each prove one or two central results, which are given at the beginning of the chapter (except for Chapters 10 and 11 which deal with the experiments and give an outlook). Except for the results which will be referenced in subsequent chapters as well, it should be possible to read and understand every chapter on its own (provided the introduction has been read). In order to aid the reader an index of notation is added in the appendix.

Part I

Random Hypergraphs and their Giant Component

Chapter 1

Introduction

While studying a similarity network of molecules for structural peculiarities we observed the striking fact that the evolution of its largest component behaves rather oddly (see Figure 1 in the preface) in that it exhibits a slow growth of the largest component before a sudden jump to (almost complete) connectivity. One idea for the underlying reasons of this behaviour was that the vertices are added in clusters and not individually thereby letting the largest component grow at moderate speed. This idea leads directly to the model of random hypergraphs where the insertion of an hyperedge containing d vertices is (from the viewpoint of connectivity) equivalent to adding all pairwise connections among the d vertices.

Although the component structure and the connectedness of a random graph belong to the most thoroughly studied subjects in the field, less is known concerning random hypergraphs. One of our goals is to give asymptotic results for a number of properties related to connectivity (for instance the asymptotic number of connected hypergraphs with a given number of edges and vertices). The most important tool to achieve this goal is the local limit theorem for the order of the giant component which we prove in Chapter 3.

Let $H = (V, E)$ denote a d -uniform hypergraph with a set V of vertices and a set E of edges, which are subsets of V of cardinality d . A vertex w is *reachable in H* from a vertex v if either $v = w$ or there is a sequence e_1, \dots, e_k of edges such that $v \in e_1$, $w \in e_k$, and $e_i \cap e_{i+1} \neq \emptyset$ for $i = 1, \dots, k-1$. Of course, reachability in H is an equivalence relation. The equivalence classes are the *components* of H , and H is *connected* if there is only one component.

Throughout this part, we let $V = \{1, \dots, n\}$ be a set of n vertices. Moreover, if $2 \leq d$ is a fixed integer and $0 \leq p = p(n) \leq 1$ is sequence of edge probabilities, then we let $H_d(n, p)$ signify a random d -uniform hypergraph with vertex set V in which each of the $\binom{n}{d}$ possible edges is present with probability p independently. We say that $H_d(n, p)$ enjoys some property \mathcal{P} *asymptotically almost surely* (a.a.s.) if the probability that $H_d(n, p)$ has \mathcal{P} tends to 1 as $n \rightarrow \infty$. If $d = 2$, then the $H_d(n, p)$ model is identical with the well-known $G_{n, p}$ model of random graphs. We will also prove results concerning a different model for random hypergraphs ($H_d(n, m)$), where the hypergraph is chosen

uniformly at random among all d -uniform hypergraphs with n vertices and m edges. In the case of $m = p\binom{n}{d}$ both models are often equivalent (see for instance [Janson et al., 2000, Section 1.4]).

The giant component.

In their seminal work on random graphs, Erdős and Rényi [1960] proved that the number of vertices in the largest component of $G_{n,p}$ undergoes a *phase transition* as $np \sim 1$. They showed that if $np < 1 - \varepsilon$ for an arbitrarily small $\varepsilon > 0$ that remains fixed as $n \rightarrow \infty$, then all components of $G_{n,p}$ consist of $O(\ln n)$ vertices. By contrast, if $np > 1 + \varepsilon$, then $G_{n,p}$ has one *giant* component on a linear number $\Omega(n)$ of vertices, while all other components contain only $O(\ln n)$ vertices. In fact, in the case $1 + \varepsilon < c = (n-1)p = O(1)$ Erdős and Rényi also estimated the order (i.e., the number of vertices) of the giant component: let $\mathcal{N}(G_{n,p})$ signify the maximum order of a component of $G_{n,p}$. Then

$$n^{-1}\mathcal{N}(G_{n,p}) \text{ converges in distribution to the constant } 1 - \rho, \quad (1.1)$$

where $0 < \rho < 1$ is the unique solution to the transcendental equation $\rho = \exp(c(\rho - 1))$.

A corresponding result was established by Schmidt-Pruzan and Shamir [1985] for random hypergraphs $H_d(n, p)$. They showed that a random hypergraph $H_d(n, p)$ consists of components of order $O(\ln n)$ if $(d-1)\binom{n-1}{d-1}p < 1 - \varepsilon$, whereas $H_d(n, p)$ has a unique large (the *giant*) component on $\Omega(n)$ vertices a.a.s. if $(d-1)\binom{n-1}{d-1}p > 1 + \varepsilon$. Furthermore, Coja-Oghlan et al. [2006] established a result similar to (1.1), showing that in the case $c := (d-1)\binom{n-1}{d-1}p > 1 + \varepsilon$ the order of the giant component is $(1 - \rho)n + o(n)$ a.a.s., where $0 < \rho < 1$ is the solution to the transcendental equation

$$\rho = \exp(c(\rho^{d-1} - 1)). \quad (1.2)$$

Since the pioneering work of Erdős and Rényi, the component structure of random graphs has been a central theme in the theory of random discrete structures. In the present work, we contribute to this theme by analysing the order (number of vertices, $\mathcal{N}(H_d(n, p))$) and the size (number of edges, $\mathcal{M}(H_d(n, p))$) of the giant component in greater detail. More precisely, establishing central and local limit theorems for $\mathcal{N}, \mathcal{M}(H_d(n, p))$, we determine the asymptotic joint distribution of $\mathcal{N}, \mathcal{M}(H_d(n, p))$ and $\mathcal{N}, \mathcal{M}(H_d(n, m))$ precisely. Though such limit theorems are known in the case of graphs (i.e., $d = 2$; cf. also the related work below), they are new in the case of d -uniform hypergraphs for $d > 2$. This is also due to the fact that none of the arguments for the graph case is directly applicable to the case of hypergraphs (for $d > 2$). Furthermore, we present a new, purely probabilistic proof of the central and local limit theorems, which, in contrast to prior work, does not rely on involved enumerative techniques or on analysing the probability that a random graph $G_{n,p}$ is connected.

These results together with the fact that the giant component is a uniform random connected hypergraph (conditioned on its order and size) will enable us to give asymptotic formulas for the probability of connectedness in the $H_d(n, p)$ and the $H_d(n, m)$

model which in turn allows to precisely estimate the asymptotic number of connected hypergraphs.

We believe that the techniques used are interesting not only for $H_d(n, p)$ with $d > 2$, but also in the case of random graphs $G_{n, p}$ because our approach leads to the first unified solution to the problems mentioned for $G_{n, p}$ as well.

1.1 Related work

Graphs.

Bender et al. [1990] were the first to compute the asymptotic probability that a random graph $G_{n, m}$ is connected for *any* ratio m/n . Although they employ a probabilistic result from Łuczak [1990] to simplify their arguments, their proof is based on enumerative considerations. Using their formula for the connectivity probability of $G_{n, m}$, Bender et al. [1992] inferred the probability that $G_{n, p}$ is connected as well as a central limit theorem for the number of edges of $G_{n, p}$ given connectedness. Moreover, it is possible (though somewhat technical) to derive local limit theorems for $\mathcal{N}, \mathcal{M}(G_{n, m})$ and $\mathcal{N}, \mathcal{M}(G_{n, p})$ from the main result of Bender et al. [1990]. In fact, Pittel and Wormald [2003, 2005] recently used enumerative arguments to rederive an improved version of the main result of Bender et al. [1990] and to obtain a local limit theorem that in addition to \mathcal{N} and \mathcal{M} also includes the order and size of the 2-core of $G_{n, m}$ or $G_{n, p}$. In summary, in Bender et al. [1990, 1992], Pittel and Wormald [2003, 2005] enumerative results on the number of connected graphs a given order and size are used to infer the distribution of $\mathcal{N}, \mathcal{M}(G_{n, p})$ and $\mathcal{N}, \mathcal{M}(G_{n, m})$. By contrast, in the present work we use the converse approach: employing probabilistic methods, we first determine the distribution of $\mathcal{N}, \mathcal{M}(G_{n, p})$ and $\mathcal{N}, \mathcal{M}(G_{n, m})$, and from this we derive the number of connected graphs with given order and size.

The asymptotic probability that $G_{n, p}$ is connected was first computed by Stepanov [1970]. He also obtains a local limit theorem for $\mathcal{N}(G_{n, p})$ (but his methods do not yield the distribution of $\mathcal{N}(G_{n, p})$ and $\mathcal{M}(G_{n, p})$). Moreover, using his result on the joint distribution of the numbers of trees of given sizes outside the giant component, Pittel [1990] derived central limit theorems for $\mathcal{N}(G_{n, p})$ and $\mathcal{N}(G_{n, m})$; the arguments in both Pittel [1990], Stepanov [1970] are of an enumerative/analytic nature.

Furthermore, a few authors have applied probabilistic arguments to problems related to the present work. For instance, O'Connell [1998] employed the theory of large deviations in order to estimate the probability that $G_{n, p}$ is connected up to a factor $\exp(o(n))$. While this result is significantly less precise than Stepanov's, O'Connell's proof is simpler. In addition, Barraez et al. [2000] exploited the analogy between the component structure of $G_{n, p}$ and branching processes to derive a central limit theorem for the joint distribution of $\mathcal{N}(G_{n, p})$ and the *total* number of edges in $G_{n, p}$; however, their techniques do not yield a *local* limit theorem. Finally, van der Hofstad and Spencer [2005] used an elegant refinement of the branching process argument to rederive the formula of Bender et al. [1990] for the number of connected graphs.

Hypergraphs.

In contrast to the case of graphs ($d = 2$), little is known about the phase transition and the connectivity probability of random d -uniform hypergraphs with $d > 2$. In fact, to the best of our knowledge the arguments used in all of the aforementioned papers do not extend to the case $d > 2$.

Karoński and Łuczak [1997] derived an asymptotic formula for the number of connected d -uniform hypergraphs of order n and size $m = n/(d-1) + o(\ln n / \ln \ln n)$ via combinatorial techniques. Since the minimum number of edges necessary for connectivity is $n/(d-1)$, this result addresses *sparsely* connected hypergraphs. Using this result, Karoński and Łuczak [2002] investigated the phase transition of $H_d(n, p)$. They established (among other things) a local limit theorem for $\mathcal{N}(H_d(n, m))$ for $m = n/d(d-1) + l$ and $1 \ll \frac{l^3}{n^2} \leq \frac{\ln n}{\ln \ln n}$ which is similar to $H_d(n, p)$ at the regime $\binom{n-1}{d-1}p = (d-1)^{-1} + \omega$, where $n^{-1/3} \ll \omega = \omega(n) \ll n^{-1/3} \ln n / \ln \ln n$. The counting result was extended by Andriamampianina and Ravelomanana [2005], Ravelomanana and Rijamamy [2005] to the regime $l = o(n^{1/3})$ ($\omega = o(n^{-2/3})$) respectively). Note that all of these results either deal with *sparsely* connected hypergraphs (i.e., $m = (d-1)^{-1}n + o(n)$), or with the *early* supercritical phase (i.e., $m = \binom{n}{d}p = (d-1)^{-1}n + o(n)$). By contrast, our results concern connected hypergraphs with $m = (d-1)^{-1}n + \Omega(n)$ edges and the component structure of random hypergraphs $H_d(n, m)$ or $H_d(n, p)$ with $m = \binom{n}{d}p = (d-1)^{-1}n + o(n)$. Thus, our results and those of Andriamampianina and Ravelomanana [2005], Karoński and Łuczak [1997, 2002], Ravelomanana and Rijamamy [2005] are complementary.

The regime of m and p that we deal with in the present work was previously studied by Coja-Oghlan et al. [2006] using probabilistic arguments. Setting up an analogy between a certain branching process and the component structure of $H_d(n, p)$, Coja-Oghlan, Moore, and Sanwalani computed the expected order and size of the largest component of $H_d(n, p)$ along with the variance of $\mathcal{N}(H_d(n, p))$. Furthermore, they computed the probability that $H_d(n, m)$ or $H_d(n, p)$ is connected *up to a constant factor*, and estimated the *expected* number of edges of $H_d(n, p)$ given connectivity. Note that Theorems 5.1, 5.2, and 5.3 enhance these results considerably, as they yield tight asymptotics for the connectivity probability, respectively the precise limiting distribution of the number of edges given connectivity.

1.2 Techniques and outline.

The aforementioned work of Andriamampianina and Ravelomanana [2005], Karoński and Łuczak [1997, 2002] on the giant component for random hypergraphs relies on enumerative techniques to a significant extent; for the basis Andriamampianina and Ravelomanana [2005], Karoński and Łuczak [1997, 2002] are results on the asymptotic number of connected hypergraphs with a given number of vertices and edges. By contrast, in the present work we employ neither enumerative techniques nor results, but rely solely on probabilistic methods. Our proof methods are also quite different from Stepanov [1970], who first estimates the asymptotic probability that a random graph $G_{n,p}$ is connected

in order to determine the distribution of $\mathcal{N}(G_{n,p})$. By contrast, in the present work we prove the local limit theorem for $\mathcal{N}(H_d(n,p))$ directly, thereby obtaining “en passant” a new proof for the local limit theorem for random graphs $G_{n,p}$, which may be of independent interest. Besides, the local limit theorem can be used to compute the asymptotic probability that $G_{n,p}$ or, more generally, $H_d(n,p)$ is connected, or to compute the asymptotic number of connected hypergraphs with a given number of vertices and edges (cf. Chapter 5). Hence, the general approach taken in the present work is actually converse to the prior ones Andriamampianina and Ravelomanana [2005], Karoński and Łuczak [1997, 2002], Stepanov [1970].

The proof of Theorem 2.1 makes use of *Stein’s method*, which is a general technique for proving central limit theorems (Stein [1970]). Roughly speaking, Stein’s result implies that a sum of a family of dependent random variables converges to the normal distribution if one can bound the correlations within any constant-sized subfamily sufficiently well. The method was used by Barbour et al. [1989] in order to prove that in a random graph $G_{n,p}$, e.g., the number of tree components of a given (bounded) size is asymptotically normal. To establish Theorem 2.1, we extend their techniques in two ways.

- Instead of dealing with the number of vertices in trees of a given size, we apply Stein’s method to the *total* number $n - \mathcal{N}(H_d(n,p))$ of vertices outside of the giant component; this essentially means that we need to sum over all possible tree sizes up to about $\ln n$.
- Since we are dealing with hypergraphs rather than graphs, we are facing a somewhat more complex situation than Barbour et al. [1989], because the fact that an edge may involve an arbitrary number d of vertices yields additional dependencies.

The main contribution of the first part of this thesis is the proof of Theorem 3.1. To this end, we think of the edges of $H_d(n,p)$ as being added in two “portions”. More precisely, we first include each possible edge with probability $p_1 = (1 - \varepsilon)p$ independently, where $\varepsilon > 0$ is small but independent of n (and denote the resulting random hypergraph by H_1); by Theorem 2.1, the order $\mathcal{N}(H_1)$ of the largest component of H_1 is asymptotically normal. Then, we add each possible edge that is not present in H_1 with a small probability $p_2 \sim \varepsilon p$ and investigate closely how these additional random edges attach further vertices to the largest component of H_1 . Denoting the number of these “attached” vertices by \mathcal{S} , we will show that the conditional distribution of \mathcal{S} *given the value of $\mathcal{N}(H_1)$* satisfies a local limit theorem. Since p_1 and p_2 are chosen such that each edge is present with probability p after the second portion of edges has been added, this yields the desired result on $\mathcal{N}(H_d(n,p))$.

The analysis of the conditional distribution of \mathcal{S} involves proving that \mathcal{S} is asymptotically normal. To show this, we employ Stein’s method once more. In addition, in order to show that \mathcal{S} satisfies a *local* limit theorem, we prove that the number of isolated vertices of H_1 that get attached to the largest component of H_1 by the second portion of random edges is binomially distributed. Since the binomial distribution satisfies a local limit theorem, we thus obtain a local limit theorem for \mathcal{S} .

Our proof of Theorem 3.1 makes use of some results on the component structure of $H_d(n, p)$ derived in Coja-Oghlan et al. [2006]. For instance, we employ the results on the expectation and the variance of $\mathcal{N}(H_d(n, p))$ from that paper. Furthermore, the analysis of \mathcal{S} given in the present work is a considerable extension of the argument used in Coja-Oghlan et al. [2006] in order to estimate the probability that $H_d(n, p)$ is connected up to a constant factor.

To prove Theorems 4.1 and 4.3, we build upon a qualitative result on the connected components of $H_d(n, p)$ from Coja-Oghlan et al. [2006] (Theorems 1.2 and 3.1, cf. Section 1.3). The proofs of these ingredients solely rely on probabilistic reasoning (namely, branching processes and Stein’s method for proving convergence to a Gaussian).

In Section 4.2 we show that (somewhat surprisingly) the *univariate* local limit theorem for $\mathcal{N}(H_d(n, p))$ can be converted into a *bivariate* local limit theorem for $\mathcal{N}(H_d(n, m))$ and $\mathcal{M}(H_d(n, m))$. To this end, we observe that the local limit theorem for $\mathcal{N}(H_d(n, p))$ implies a bivariate local limit theorem for the joint distribution of $\mathcal{N}(H_d(n, p))$ and the number $\mathcal{M}(H_d(n, p))$ of edges *outside* the largest component. Then, we will set up a relationship between the joint distribution of $\mathcal{N}, \bar{\mathcal{M}}(H_d(n, p))$ and that of $\mathcal{N}, \bar{\mathcal{M}}(H_d(n, m))$. Since we already know the distribution of $\mathcal{N}, \bar{\mathcal{M}}(H_d(n, p))$, we can infer the joint distribution of $\mathcal{N}, \bar{\mathcal{M}}(H_d(n, m))$ via Fourier analysis. As in $H_d(n, m)$ the *total* number of edges is fixed (namely, m), we have $\bar{\mathcal{M}}(H_d(n, m)) = m - \mathcal{M}(H_d(n, m))$. Hence, we obtain a local limit theorem for the joint distribution of $\mathcal{N}, \mathcal{M}(H_d(n, m))$, i.e., Theorem 4.3. Finally, Theorem 4.3 easily implies Theorem 4.1. We actually consider this Fourier analytic approach for proving the bivariate local limit theorems the main contribution of the present work.

Furthermore, in Section 5.4 we derive Theorem 5.1 from Theorem 4.1. The basic reason why this is possible is that *given* that the largest component of $H_d(n, p)$ has order ν and size μ , this component is a uniformly distributed random hypergraph with these parameters. Indeed, this observation was also exploited by Łuczak [1990] to estimate the number of connected graphs up to a polynomial factor, and in Coja-Oghlan et al. [2006], where an explicit relation between $C_d(\nu, \mu)$ and $\mathbb{P}[\mathcal{N}(H_d(n, p)) = \nu \wedge \mathcal{M}(H_d(n, p)) = \mu]$ was derived. Combining this formula with Theorem 4.1, we obtain Theorem 5.1. Moreover, in Sections 5.3 and 5.5 we use similar arguments to establish Theorems 5.2 and 5.3.

The main part is organised as follows. After making some preliminaries in Section 1.3, we prove the central limit theorem for $\mathcal{N}(H_d(n, p))$ via Stein’s method in Chapter 2. We outline the proof of the Local Limit Theorem 3.1 in Section 3.2. In that section we explain in detail how $H_d(n, p)$ is generated in two “portions”. Then, in Section 3.3 we analyse the random variable \mathcal{S} , assuming the central limit theorem for \mathcal{S} . Further, Section 3.4 deals with the proof the central limit theorem for \mathcal{S} via Stein’s method reusing the arguments of Chapter 2. Chapter 4 contains the proofs of additional local limit theorems for the different random graph models and joint distributions while in Chapter 5 we apply our results to get some statements about the connectivity probability and the number of connected hypergraphs.

1.3 Preliminaries

Throughout the whole part, we let $V = \{1, \dots, n\}$. If $d \geq 2$ is an integer and $V_1, \dots, V_k \subset V$, then we let $\mathcal{E}_d(V_1, \dots, V_k)$ signify the set of all subsets $e \subset V$ of cardinality d such that $e \cap V_i \neq \emptyset$ for all i . We omit the subscript d if it is clear from the context.

If H is a hypergraph, then we let $V(H)$ denote its vertex set and $E(H)$ its edge set. We say that a set $S \subset V(H)$ is *reachable from* $T \subset V(H)$ if each vertex $s \in S$ is reachable from some vertex $t \in T$. Further, if $V(H) \subset V = \{1, \dots, n\}$, then the subsets of V can be ordered lexicographically; hence, we can define the *largest component* of H to be the lexicographically first component of order $\mathcal{N}(H)$.

We use the O -notation to express asymptotic estimates as $n \rightarrow \infty$ and abbreviate $f(n) = (1 + o(1))g(n)$ by $f(n) \sim g(n)$. Furthermore, if $f(x_1, \dots, x_k, n)$ is a function that depends not only on n but also on some further parameters x_i from domains $D_i \subset \mathbb{R}$ ($1 \leq i \leq k$), and if $g(n) \geq 0$ is another function, then we say that the estimate $f(x_1, \dots, x_k, n) = O(g(n))$ holds *uniformly in* x_1, \dots, x_k if the following is true: if \mathcal{I}_j and D_j , $\mathcal{I}_j \subset D_j$, are compact sets, then there exist numbers $C = C(\mathcal{I}_1, \dots, \mathcal{I}_k)$ and $n_0 = n_0(\mathcal{I}_1, \dots, \mathcal{I}_k)$ such that $|f(x_1, \dots, x_k, n)| \leq Cg(n)$ for all $n \geq n_0$ and $(x_1, \dots, x_k) \in \prod_{j=1}^k \mathcal{I}_j$. We define uniformity analogously for the other Landau symbols Ω , Θ , etc.

We shall make repeated use of the following *Chernoff bound* on the tails of a binomially distributed variable $X = \text{Bi}(\nu, q)$ (cf. [Janson et al., 2000, p. 26] for a proof): for any $t > 0$ we have

$$\mathbb{P}[|X - \mathbb{E}[X]| \geq t] \leq 2 \exp \left(-\frac{t^2}{2(\mathbb{E}[X] + t/3)} \right). \quad (1.3)$$

Moreover, we employ the following *local limit theorem* for the binomial distribution (cf. [Bollobás, 2001, Chapter 1]).

Proposition 1.1. *Suppose that $0 \leq p = p(n) \leq 1$ is a sequence such that $np(1-p) \rightarrow \infty$ as $n \rightarrow \infty$. Let $X = \text{Bi}(n, p)$. Then for any sequence $x = x(n)$ of integers such that $|x - np| = o(np(1-p)^{2/3})$,*

$$\mathbb{P}[X = x] \sim (2\pi np(1-p))^{-\frac{1}{2}} \exp \left(-\frac{(x - np)^2}{2p(1-p)n} \right) \quad \text{as } n \rightarrow \infty.$$

Furthermore, we use the following theorem, which summarises results from [Coja-Oghlan et al., 2006, Section 6] on the component structure of $H_d(n, p)$.

Theorem 1.2. *Let $p = c \binom{n-1}{d-1}^{-1}$.*

1. *If there is a fixed $c_0 < (d-1)^{-1}$ such that $c = c(n) \leq c_0$, then*

$$\mathbb{P}[\mathcal{N}(H_d(n, p)) \leq 3(d-1)^2(1 - (d-1)c_0)^{-2} \ln n] \geq 1 - n^{-100}.$$

2. *Suppose that $c_0 > (d-1)^{-1}$ is a constant, and that $c_0 \leq c = c(n) = o(\ln n)$ as $n \rightarrow \infty$. Then the transcendental equation (1.2) has a unique solution $0 < \rho = \rho(c) < 1$, which satisfies*

$$\rho^{d-1}c < c'_0 < (d-1)^{-1}. \quad (1.4)$$

for some number $c'_0 > 0$ that depends only on c_0 . Moreover,

$$\begin{aligned} |\mathbb{E}[\mathcal{N}(H_d(n, p))] - (1 - \rho)n| &\leq n^{o(1)}, \\ \text{Var}[\mathcal{N}(H_d(n, p))] &\sim \frac{\rho(1 - \rho + c(d - 1)(\rho - \rho^{d-1}))n}{(1 - c(d - 1)\rho^{d-1})^2}. \end{aligned}$$

Furthermore, with probability $\geq 1 - n^{-100}$ there is precisely one component of order $(1 + o(1))(1 - \rho)n$ in $H_d(n, p)$, while all other components have order $\leq \ln^2 n$. In addition,

$$\mathbb{P}\left[|\mathcal{N}(H_d(n, p)) - \mathbb{E}[\mathcal{N}(H_d(n, p))]| \geq n^{0.51}\right] \leq n^{-100}.$$

Finally, the following result on the component structure of $H_d(n, p)$ with average degree $\binom{n-1}{d-1}p < (d - 1)^{-1}$ below the threshold has been derived in [Coja-Oghlan et al., 2006, Section 6] via the theory of branching processes.

Proposition 1.3. *There exists a function $q : (0, (d - 1)^{-1}) \times (0, 1) \rightarrow \mathbb{R}_{\geq 0}$, $(\zeta, \xi) \mapsto q(\zeta, \xi) = \sum_{k=1}^{\infty} q_k(\zeta)\xi^k$ whose coefficients $\zeta \mapsto q_k(\zeta)$ are differentiable such that the following holds. Suppose that $0 \leq p = p(n) \leq 1$ is a sequence such that $0 < \binom{n-1}{d-1}p = c = c(n) < (d - 1)^{-1} - \varepsilon$ for an arbitrarily small $\varepsilon > 0$ that remains fixed as $n \rightarrow \infty$. Let $P(c, k)$ denote the probability that in $H_d(n, p)$ some fixed vertex $v \in V$ lies in a component of order k . Then*

$$P(c, k) = (1 + o(n^{-2/3}))q_k(c) \quad \text{for all } 1 \leq k \leq \ln^2 n. \quad (1.5)$$

Furthermore, for any fixed $\varepsilon > 0$ there is a number $0 < \gamma = \gamma(\varepsilon) < 1$ such that

$$q_k(c) \leq \gamma^k \quad \text{for all } 0 < c < (d - 1)^{-1} - \varepsilon. \quad (1.6)$$

Lemma 1.4.

$$\mathbb{P}[|C_v| = k] = (1 + O(n^{-1} \cdot \text{polylog } n))\mathbb{P}[T = k] \text{ for } k = O(\text{polylog } n).$$

where T denotes the stopping time of a branching process with successor distribution $(d - 1)\text{Po}(c)$ with $\text{Po}(c)$ being the Poisson distribution with mean c .

Proof. We discover the component of v via a branching process just as in Coja-Oghlan et al. [2006]. Proposition 30 in Coja-Oghlan et al. [2006] shows that the number of explored vertices in each epoch i of the branching process is a random variable Z_i^* which is dominated by another random variable Z_i' and dominates a third Z_i'' . According to Lemma 29 in Coja-Oghlan et al. [2006] the random variables T' and T'' corresponding to the stopping times of the branching processes on Z_i' and Z_i'' are distributed such that

$$\begin{aligned} \mathbb{P}[T' = k] &= (1 + O(n^{-1} \cdot \text{polylog } n))\mathbb{P}[T = k], \\ \mathbb{P}[T'' = k] &= (1 + O(n^{-1} \cdot \text{polylog } n))\mathbb{P}[T = k]. \end{aligned}$$

This proves the statement of the lemma. □

Proof of Proposition 1.3. Lemma 1.4 gives that the first $\ln^2 n$ coefficients of the power series expansion of \tilde{q} where \tilde{q} is the solution to

$$\tilde{q}(c, x) = x \exp(c(\tilde{q}(c, x)^{d-1} - 1)) \quad (1.7)$$

have property (1.5), since

$$\tilde{q} = \sum_{k=1}^{\infty} \mathbb{P}[T = k] x^k. \quad (1.8)$$

Now defining $q(c, x) = \sum_{k=1}^{\ln^2 n} q_k(c) x^k$, where $q_k = \mathbb{P}[T = k]$ we see that q is differentiable in x and it suffices to show that the q_k are differentiable in c . Using (1.8) we see that in order to calculate q_k we can set up a system of linear equations in the following way. Let q' denote the derivative of q with respect to x and define $\tilde{q}_{(0)} := \tilde{q}$ and inductively $\tilde{q}_{(i+1)} := x \tilde{q}'_{(i)}$. This gives a system of linear equations of the form $\sum_{k=1}^{\ln^2 n} k^i q_k = \tilde{q}_{(i)}(1)$ for $i \in [\ln^2 n]$. The coefficient vectors of the q_k are obviously linear independent for different i , thus there is a unique algebraic solution provided we have an algebraic expression for $\tilde{q}_{(i)}(1)$. We already know that $\tilde{q}_{(0)}(1) = \tilde{q}(c, 1) = 1$ and by computing the derivative with respect to x of both sides of (1.7) we get:

$$\tilde{q}'(c, x) = \frac{\tilde{q}(c, x)}{x(1 - c(d-1)\tilde{q}(c, x)^{d-1})} \quad (1.9)$$

and thus can easily calculate $\tilde{q}_{(1)}(1)$ and by further differentiating (1.9) get algebraic expressions for all $\tilde{q}_{(i)}(1)$.

The second statement of the proposition follows directly from Theorem 5 in Coja-Oghlan et al. [2006]. \square

We let $\mathcal{N}(H)$ signify the maximum order of a component of H . Furthermore, for all hypergraphs H we consider the vertex set $V(H)$ will consist of integers. Therefore, the subsets of $V(H)$ can be ordered lexicographically, and we call the lexicographically first component of H that has order $\mathcal{N}(H)$ the *largest component* of H . In addition, we denote by $\mathcal{M}(H)$ the size of the largest component of H .

We will consider the two models of random d -uniform hypergraphs: $H_d(n, p)$ and $H_d(n, m)$. The random hypergraph $H_d(n, p)$ has the vertex set $V = \{1, \dots, n\}$, and each of the $\binom{n}{d}$ possible edges is present with probability p independently of all others. Moreover, $H_d(n, m)$ is a uniformly distributed hypergraph with vertex set $V = \{1, \dots, n\}$ and with exactly m edges. In the case $d = 2$, the notation $G_{n,p} = H_2(n, p)$, $G_{n,m} = H_2(n, m)$ is commonly used.

1.3.1 The Phase Transition and the Giant Component

In their two pioneering papers on the theory of random graphs, Erdős and Rényi [1959, 1960] studied the component structure of the random graph $G_{n,m}$. Since then, the component structure of random discrete objects (e.g., graphs, hypergraphs, digraphs, ...) has been among the main subjects of discrete probability theory. One reason for this is

the connection to statistical physics and percolation (as “mean field models”); another reason is the impact of these considerations on computer science (e.g., due to relations to computational problems such as MAX CUT or MAX 2-SAT, Coppersmith et al. [2004]).

In their first paper Erdős and Rényi [1959] showed that if t remains fixed as $n \rightarrow \infty$ and $m = \frac{n}{2}(\ln n + t)$, then the probability that $G_{n,m}$ is connected is asymptotically $\exp(-\exp(t))$ as $n \rightarrow \infty$. Since $G_{n,m}$ is a uniformly distributed graph, this result immediately yields the asymptotic number of connected graphs of order n and size m . The relevance of this result notwithstanding, possibly the most important contribution of Erdős and Rényi [1959] is that they solved this *enumerative* problem (“how many connected graphs of order n and size m exist?”) via *probabilistic* methods (namely, the method of moments for proving convergence to a Poisson distribution).

Furthermore, Erdős and Rényi [1960] went on to study (among other things) the component structure of *sparse* random graphs with $m = O(n)$ edges. The main result is that the order $\mathcal{N}(G_{n,m})$ of the largest component undergoes a *phase transition* as $2m/n \sim 1$. Let us state actually state a more general version from Schmidt-Pruzan and Shamir [1985], which covers d -uniform hypergraphs: let either $H = H_d(n, m)$ and $c = dm/n$, or $H = H_d(n, p)$ and $c = \binom{n-1}{d-1}p$; we refer to c as the *average degree* of H . Then the result is that

- if $c < (d-1)^{-1} - \varepsilon$ for an arbitrarily small but fixed $\varepsilon > 0$, then $\mathcal{N}(G_{n,m}) = O(\ln n)$ a.a.s.
- By contrast, if $c > (d-1)^{-1} + \varepsilon$, then $G_{n,m}$ features a unique component of order $\Omega(n)$ a.a.s., which is called the *giant component*. More precisely, $\mathcal{N}(H) = (1-\rho)n + o(n)$ a.a.s. where ρ is the unique solution to the transcendental equation (1.2) that lies strictly between 0 and 1. Furthermore, the second largest component has order $O(\ln n)$.

Chapter 2

A Central Limit Theorem for the Number of Vertices

2.1 Results

In terms of limit theorems, (1.1) provides a *strong law of large numbers* for $\mathcal{N}(G_{n,p})$, i.e., it yields the probable value of $\mathcal{N}(G_{n,p})$ up to fluctuations of order $o(n)$. Thus, a natural question is if we can characterise the distribution of $\mathcal{N}(G_{n,p})$ (or $\mathcal{N}(H_d(n,p))$) more precisely; for instance, is it true that $\mathcal{N}(G_{n,p})$ “converges to the normal distribution” in some sense? Our first result, which we will prove in this chapter, shows that this is indeed the case.

Theorem 2.1. *Let $\mathcal{J} \subset ((d-1)^{-1}, \infty)$ be a compact interval, and let $0 \leq p = p(n) \leq 1$ be a sequence such that $c = c(n) = \binom{n-1}{d-1}p \in \mathcal{J}$ for all n . Furthermore, let $0 < \rho = \rho(n) < 1$ be the unique solution to (1.2), and set*

$$\sigma^2 = \sigma(n)^2 = \frac{\rho(1 - \rho + c(d-1)(\rho - \rho^{d-1}))n}{(1 - c(d-1)\rho^{d-1})^2}. \quad (2.1)$$

Then $\sigma^{-1}(\mathcal{N}(H_d(n,p)) - (1 - \rho)n)$ converges in distribution to the standard normal distribution.

Theorem 2.1 provides a *central limit theorem* for $\mathcal{N}(H_d(n,p))$; it shows that for any fixed numbers $a < b$

$$\lim_{n \rightarrow \infty} \mathbb{P} \left[a \leq \frac{\mathcal{N}(H_d(n,p)) - (1 - \rho)n}{\sigma} \leq b \right] = (2\pi)^{-\frac{1}{2}} \int_a^b \exp(-t^2/2) dt \quad (2.2)$$

(provided that the sequence $p = p(n)$ satisfies the above assumptions).

In this chapter we will use Stein’s Method to prove Theorem 2.1 saying that $\mathcal{N}(H_d(n,p))$ tends (after suitable normalisation) in distribution to the normal distribution. We will do so in a more general setting which will allow us to prove Lemma 3.10 using the same method. First we will discuss the result by Barbour et al. [1989] and how to apply it to

random hypergraphs, which yields some conditions the random variables have to fulfil. Then we show in Lemma 2.6 that the random variables corresponding to $\mathcal{N}(H_d(n, p))$ do indeed comply to the conditions and last but not least a quite technical part will show how to derive the limiting distribution from the conditions.

Instead of analysing the distribution of the number of vertices in the giant component directly we will rather count the number of vertices in isolated trees of up to polylogarithmic order, since it is well known, that the number of vertices which belong neither to the giant nor to an isolated tree is $O(1)$ (cf. [Janson et al., 2000, Chapter 5]).

The main result from Barbour et al. [1989] about Stein's method is the following.

Theorem 2.2. *Let W be random variable which gets decomposed using finite index sets I and $K_i \subseteq I$, $i \in I$ and sets of square integrable random variables $X_i, W_i, Z_i, Z_{ik}, W_{ik}, V_{ik}$ in the following way:*

$$W = \sum_{i \in I} X_i, \quad (2.3)$$

$$\mathbb{E}[X_i] = 0, i \in I, \quad \mathbb{E}[W^2] = 1, \quad (2.4)$$

$$W = W_i + Z_i, i \in I, \quad \text{where } W_i \text{ is independent of } X_i, \quad (2.5)$$

$$Z_i = \sum_{k \in K_i} Z_{ik}, i \in I, \quad (2.6)$$

$$W_i = W_{ik} + Z_{ik}, i \in I, k \in K_i \quad \text{where } W_{ik} \text{ is independent of the pair } (X_i, Z_{ik}). \quad (2.7)$$

Then

$$d_1\left(\frac{W - \mathbb{E}[W]}{\sqrt{\text{Var}[W]}}, \phi_{0,1}\right) = O(\delta)$$

where

$$d_1(A, B) := \sup_h \left\{ \frac{|\mathbb{E}[h(A)] - \mathbb{E}[h(B)]|}{\sup_{x \in \mathbb{R}} |h(x)| + \sup_{x \in \mathbb{R}} |h'(x)|} \right\}$$

and

$$\delta := \frac{1}{2} \sum_{i \in I} \mathbb{E}[|X_i|Z_i^2] + \sum_{i \in I} \sum_{k \in K_i} (\mathbb{E}[|X_i Z_{ik} V_{ik}|] + \mathbb{E}[|X_i Z_{ik}|] \mathbb{E}[|Z_i + V_{ik}|])$$

2.2 Stein's Method for Random Hypergraphs

Let \mathcal{E} be the set of all subsets of size d of $V = \{1, \dots, n\}$, and let \mathcal{H} be the power set of \mathcal{E} . Moreover, let $0 \leq p_e \leq 1$ for each $e \in \mathcal{E}$, and define a probability distribution on \mathcal{H} by letting $\mathbb{P}[H] = \prod_{e \in H} p_e \cdot \prod_{e \in \mathcal{E} \setminus H} 1 - p_e$. That is $H \in \mathcal{H}$ can be considered a random hypergraph with "individual" edge probabilities.

Furthermore, let \mathcal{A} be a family of subsets of V , and let $(Y_\alpha)_{\alpha \in \mathcal{A}}$ be a family of random variables. Remember that for $Q \subset V$ we set $\mathcal{E}(Q) = \{e \in \mathcal{E} : e \cap Q \neq \emptyset\}$. We say that Y_α is *feasible* if the following holds.

For any two elements $H, H' \in \mathcal{H}$ such that $H \cap \mathcal{E}(\alpha) = H' \cap \mathcal{E}(\alpha)$ we have $Y_\alpha(H) = Y_\alpha(H')$.

That means Y_α is feasible if its value depends only on edges having at least one endpoint in α . In addition, set $Y_\alpha^S(H) = Y_\alpha(H \setminus \mathcal{E}(S))$ ($H \in \mathcal{H}$, $\alpha \in \mathcal{A}$, $S \subset V$, $S \cap \alpha = \emptyset$). Thus $Y_\alpha^S(H)$ is the value of Y_α after removing all edges incident with S . We define

$$Y = \sum_{\alpha \in \mathcal{A}} Y_\alpha, \quad \mu_\alpha = \mathbb{E}[Y_\alpha], \quad \sigma^2 = \text{Var}[Y], \quad X_\alpha = (Y_\alpha - \mu_\alpha)/\sigma \quad (2.8)$$

$$Z_\alpha = \sum_{\beta \in \mathcal{A}} Z_{\alpha\beta}, \quad \text{where } Z_{\alpha\beta} = \sigma^{-1} \times \begin{cases} Y_\beta & \text{if } \alpha \cap \beta \neq \emptyset, \\ Y_\beta - Y_\beta^\alpha & \text{if } \alpha \cap \beta = \emptyset, \end{cases} \quad (2.9)$$

$$V_{\alpha\beta} = \sum_{\substack{\gamma: \beta \cap \gamma \neq \emptyset \\ \wedge \alpha \cap \gamma = \emptyset}} Y_\gamma^\alpha / \sigma + \sum_{\substack{\gamma: \beta \cap \gamma = \emptyset \\ \wedge \alpha \cap \gamma = \emptyset}} (Y_\gamma^\alpha - Y_\gamma^{\alpha \cup \beta}) / \sigma, \quad \text{and} \quad (2.10)$$

$$\delta = \sum_{\alpha \in \mathcal{A}} \mathbb{E}[|X_\alpha|^2] + \sum_{\alpha, \beta \in \mathcal{A}} (\mathbb{E}[|X_\alpha Z_{\alpha\beta} V_{\alpha\beta}|] + \mathbb{E}[|X_\alpha Z_{\alpha\beta}|] \mathbb{E}[|Z_\alpha + V_{\alpha\beta}|]). \quad (2.11)$$

The following theorem was proven for graphs (i.e. $d = 2$) in Barbour et al. [1989].

Theorem 2.3. *Suppose that all Y_α are feasible. Then Y tends to the normal distribution with mean $\mathbb{E}[Y]$ and variance $\text{Var}[Y]$ in the following sense:*

$$d_1\left(\frac{Y - \mathbb{E}[Y]}{\sqrt{\text{Var}[Y]}}, \phi_{0,1}\right) < \delta.$$

Proof. Although Barbour et al. [1989] state this result for graphs only, their argumentation carries over to the case of $H_d(n, p)$ without any essential modifications.

In order to use Theorem 2.2 we first identify the variables. The index sets $I = \mathcal{A}$ and $K_\alpha = \mathcal{A}$. W is implicitly given as the sum of the X_α . The definition of X_α (2.8) also shows that (2.4) holds. Let

$$W_\alpha := \sum_{\alpha \cap \beta = \emptyset} X_\beta^\alpha - \mathbb{E}[Z_\alpha], \quad X_\beta^\alpha := \sigma^{-1}(Y_\beta^\alpha - \mathbb{E}[Y_\beta^\alpha])$$

The fact that $W = W_\alpha + Z_\alpha$ comes directly from the corresponding definitions and W_α is independent of X_α for every α , since W_α depends on X_β^α only and the edges deciding the value X_α are all removed when looking at the value of X_β^α .

With

$$W_{\alpha\beta} := \sum_{\gamma \cap (\alpha \cup \beta) = \emptyset} X_\gamma^{\alpha \cup \beta} - \mathbb{E}[V_{\alpha\beta}] - \mathbb{E}[Z_\alpha]$$

the same is true for the independence of $W_{\alpha\beta}$ and $(X_\alpha, Z_{\alpha\beta})$, since $(X_\alpha, Z_{\alpha\beta})$ depends on edges which were removed when calculating $W_{\alpha\beta}$. $W_\alpha = W_{\alpha\beta} + Z_{\alpha\beta}$ follows again directly from the definitions. Thus all assumptions of Theorem 2.2 are fulfilled and the statement follows. \square

Now the following lemma states that a number of conditions on the expectations of the product of up to three random variables Y_α^S will suffice for $\delta = o(1)$.

Lemma 2.4. *Let $k = O(\text{polylog } n)$ and let $(Y_\alpha)_{\alpha \in \mathcal{A}}$ be a feasible family such that $0 \leq Y_\alpha \leq k$ for all $\alpha \in \mathcal{A}$. If the following six conditions are satisfied, then $\delta = O(n^{-1/2})$.*

Y1. *We have $\mathbb{E}[Y], \text{Var}[Y] = \Theta(n)$, and $\sum_{\beta \in \mathcal{A}: \beta \cap \alpha \neq \emptyset} \mu_\beta = O(\mathbb{E}[Y]/n \cdot \text{polylog } n) = O(\text{polylog } n)$ for any $\alpha \in \mathcal{A}$.*

Y2. *Let α, β, γ be distinct elements of \mathcal{A} . Then*

$$Y_\alpha(Y_\beta - Y_\beta^\alpha)Y_\beta^\alpha = 0 \quad \text{if } \alpha \cap \beta = \emptyset, \quad (2.12)$$

$$Y_\alpha Y_\beta = 0 \quad \text{if } \alpha \cap \beta \neq \emptyset, \quad (2.13)$$

$$(Y_\beta - Y_\beta^\alpha)Y_\gamma^\alpha = (Y_\beta - Y_\beta^\alpha)Y_\gamma = 0 \quad \text{if } \alpha \cap \beta = \alpha \cap \gamma = \emptyset \neq \beta \cap \gamma. \quad (2.14)$$

Y3. *For all α, β we have $\sum_{\gamma: \gamma \cap \beta \neq \emptyset, \gamma \cap \alpha = \emptyset} \mathbb{E}[Y_\beta Y_\gamma^\alpha] \leq k^2 \mu_\beta$.*

Y4. *If $\alpha, \beta \in \mathcal{A}$ are disjoint, then*

$$\mathbb{E}[Y_\alpha Y_\beta] = O(\mu_\alpha \mu_\beta \cdot \text{polylog } n), \quad (2.15)$$

$$\mathbb{E}[|Y_\beta - Y_\beta^\alpha|] = O\left(\frac{\mu_\beta}{n} \cdot \text{polylog } n\right), \quad (2.16)$$

$$\mathbb{E}[Y_\alpha | Y_\beta - Y_\beta^\alpha|] = O\left(\frac{\mu_\alpha \mu_\beta}{n} \cdot \text{polylog } n\right). \quad (2.17)$$

Y5. *If $\alpha, \beta, \gamma \in \mathcal{A}$ are pairwise disjoint, then*

$$\mathbb{E}[Y_\beta | Y_\gamma^\alpha - Y_\gamma^{\alpha \cup \beta}|] = O\left(\frac{\mu_\beta \mu_\gamma}{n} \cdot \text{polylog } n\right), \quad (2.18)$$

$$\mathbb{E}[|Y_\beta - Y_\beta^\alpha| \cdot |Y_\gamma^\alpha - Y_\gamma^{\alpha \cup \beta}|] = O\left(\frac{\mu_\beta \mu_\gamma}{n^2} \cdot \text{polylog } n\right), \quad (2.19)$$

$$\mathbb{E}[Y_\alpha | Y_\beta - Y_\beta^\alpha| \cdot |Y_\gamma^\alpha - Y_\gamma^{\alpha \cup \beta}|] = O\left(\frac{\mu_\alpha \mu_\beta \mu_\gamma}{n^2} \cdot \text{polylog } n\right), \quad (2.20)$$

$$\mathbb{E}[Y_\alpha | Y_\beta - Y_\beta^\alpha| \cdot |Y_\gamma - Y_\gamma^\alpha|] = O\left(\frac{\mu_\alpha \mu_\beta \mu_\gamma}{n^2} \cdot \text{polylog } n\right), \quad (2.21)$$

$$\mathbb{E}[|(Y_\beta - Y_\beta^\alpha)(Y_\gamma - Y_\gamma^\alpha)|] = O\left(\frac{\mu_\alpha \mu_\beta}{n^2} \cdot \text{polylog } n\right). \quad (2.22)$$

Y6. *If $\alpha, \beta, \gamma \in \mathcal{A}$ satisfy $\alpha \cap \beta = \alpha \cap \gamma = \emptyset$, then*

$$\mathbb{E}[|Y_\alpha^\beta - Y_\alpha^{\beta \cup \gamma}|] = O\left(\frac{\mu_\gamma}{n} \cdot \text{polylog } n\right). \quad (2.23)$$

Before we prove the lemma a short corollary will show how to derive a total variation distance from the d_1 distance resulting from Theorem 2.2.

Corollary 2.5. *Let Y be as in Lemma 2.4 with $d_1\left(\frac{Y - \mathbb{E}[Y]}{\sqrt{\text{Var}[Y]}}, \phi_{0,1}\right) = O(\text{Var}[Y]^{-1/2})$. Then we have for all constants a and b*

$$\mathbb{P}\left[\frac{Y - \mathbb{E}[Y]}{\sqrt{\text{Var}[Y]}} \in [a, b]\right] = (1 + O(\text{Var}[Y]^{-1/4}))\mathbb{P}[\phi_{0,1} \in [a, b]]$$

Proof. Let $h(x)$ be a differentiable function with $0 \leq h(x) \leq 1$ such that for all $x \in [a, b]$ we have $h(x) = 1$ and for $x < a - \gamma$ or $x > b + \gamma$ have $h(x) = 0$ for γ still to be defined. Furthermore let $g(x)$ be a differentiable function with $0 \leq g(x) \leq 1$ such that for all $x \notin [a, b]$ we have $g(x) = 0$ and for $x > a + \gamma$ or $x < b - \gamma$ have $g(x) = 1$. It is possible to construct such h and g such that $\sup_x |h'(x)| = \sup_x |g'(x)| < 2/\gamma$ by inserting appropriately shaped functions in the intervals of size γ .

From the definition of d_1 and the assumption on Y we know that

$$\begin{aligned} \left| \mathbb{E}\left[h\left(\frac{Y - \mathbb{E}[Y]}{\sqrt{\text{Var}[Y]}}\right)\right] - \mathbb{E}[h(\phi_{0,1})] \right| &< \left(\sup_x |h(x)| + \sup_x |h'(x)| \right) O(\text{Var}[Y]^{-1/2}) \\ &= O\left(\frac{\text{Var}[Y]^{-1/2}}{\gamma}\right) \end{aligned} \quad (2.24)$$

and

$$\begin{aligned} \left| \mathbb{E}\left[g\left(\frac{Y - \mathbb{E}[Y]}{\sqrt{\text{Var}[Y]}}\right)\right] - \mathbb{E}[g(\phi_{0,1})] \right| &< \left(\sup_x |g(x)| + \sup_x |g'(x)| \right) O(\text{Var}[Y]^{-1/2}) \\ &= O\left(\frac{\text{Var}[Y]^{-1/2}}{\gamma}\right). \end{aligned} \quad (2.25)$$

Since h and g differ only in an interval of size γ from the sharp $(0, 1)$ -function and the probability of being in an interval of size γ is always $O(\gamma)$ for a normal distribution it is clear that

$$|\mathbb{E}[h(\phi_{0,1})] - \mathbb{P}[\phi_{0,1} \in [a, b]]| = O(\gamma) \quad (2.26)$$

and

$$|\mathbb{E}[g(\phi_{0,1})] - \mathbb{P}[\phi_{0,1} \in [a, b]]| = O(\gamma). \quad (2.27)$$

Then we have (by the way we have chosen h):

$$\begin{aligned} \mathbb{P}\left[\frac{Y - \mathbb{E}[Y]}{\sqrt{\text{Var}[Y]}} \in [a, b]\right] &\leq \mathbb{E}\left[h\left(\frac{Y - \mathbb{E}[Y]}{\sqrt{\text{Var}[Y]}}\right)\right] \\ &\stackrel{(2.24)}{\leq} \mathbb{E}[h(\phi_{0,1})] + O\left(\frac{\text{Var}[Y]^{-1/2}}{\gamma}\right) \\ &\stackrel{(2.26)}{\leq} \mathbb{P}[\phi_{0,1} \in [a, b]] + O\left(\frac{\text{Var}[Y]^{-1/2}}{\gamma}\right) + O(\gamma) \end{aligned}$$

and correspondingly for g :

$$\begin{aligned}
 \mathbb{P} \left[\frac{Y - \mathbb{E}[Y]}{\sqrt{\text{Var}[Y]}} \in [a, b] \right] &\geq \mathbb{E} \left[g \left(\frac{Y - \mathbb{E}[Y]}{\sqrt{\text{Var}[Y]}} \right) \right] \\
 &\stackrel{(2.25)}{\geq} \mathbb{E}[g(\phi_{0,1})] + O \left(\frac{\text{Var}[Y]^{-1/2}}{\gamma} \right) \\
 &\stackrel{(2.27)}{\geq} \mathbb{P}[\phi_{0,1} \in [a, b]] + O \left(\frac{\text{Var}[Y]^{-1/2}}{\gamma} \right) + O(\gamma)
 \end{aligned}$$

By choosing $\gamma := \text{Var}[Y]^{-1/4}$ we get

$$\begin{aligned}
 \left| \mathbb{P} \left[\frac{Y - \mathbb{E}[Y]}{\sqrt{\text{Var}[Y]}} \in [a, b] \right] - \mathbb{P}[\phi_{0,1} \in [a, b]] \right| &O(\text{Var}[Y]^{-1/2}/\gamma + \gamma) \\
 &= O(\text{Var}[Y]^{-1/4})
 \end{aligned}$$

□

2.3 Conditions for the Normality of $\mathcal{N}(H_d(n, p))$

In this section we will prove the properties **Y1–Y6** defined in Lemma 2.4, thereby proving Theorem 2.1, which follows then directly from Theorem 2.3 together with Corollary 2.5.

Let $k = O(\text{polylog } n)$ and let $\mathcal{A} = \{\alpha \subset V : 1 \leq |\alpha| \leq k\}$. Moreover, for $A \subseteq V$ with $A \cap \alpha = \emptyset$ let $I_\alpha^A = 1$ if α is a component of $H \setminus \mathcal{E}(A)$, and 0 otherwise. Further, set $Y_\alpha^A = |\alpha| \cdot I_\alpha^A$. We briefly write $I_\alpha = I_\alpha^\emptyset$ and $Y_\alpha = Y_\alpha^\emptyset$. Then $(Y_\alpha)_{\alpha \in \mathcal{A}}$ is a feasible family, because whether α is a component or not only depends on the presence of edges that contain at least one vertex of α .

Let $\mathcal{C}(S)$ denote the event that the subhypergraph of H induced on $S \subset V$ is connected. If $I_\alpha = 1$, then $\mathcal{C}(\alpha)$ occurs. Moreover, H contains no edges joining α and $V \setminus \alpha$, i.e., $H \cap \mathcal{E}(\alpha, V \setminus \alpha) = \emptyset$. Since each edge occurs in H with probability p independently, we thus obtain

$$\mathbb{P}[I_\alpha = 1] = \mathbb{P}[\mathcal{C}(\alpha)](1 - p)^{|\mathcal{E}(\alpha, V \setminus \alpha)|}. \quad (2.28)$$

Furthermore, observe that

$$\forall \alpha \in \mathcal{A}, A \subset B \subset V \setminus \alpha : I_\alpha^A = 1 \rightarrow I_\alpha^B = 1. \quad (2.29)$$

Proof of Y1: The order of magnitude of $\mathbb{E}[Y]$ and $\text{Var}[Y]$ was already shown in [Coja-Oghlan et al., 2006, Theorem 5]. To see $\sum_{\beta \in \mathcal{A}: \beta \cap \alpha \neq \emptyset} \mu_\beta = O(\mathbb{E}[Y]/n)$ note that

$$\mathbb{E}[Y] = \sum_{\beta \in \mathcal{A}} \mu_\beta = \sum_{b=1}^k \sum_{\substack{\beta \in \mathcal{A} \\ |\beta|=b}} \mu_\beta = \sum_{b=1}^k \binom{n}{b} \mu_\beta$$

while

$$\sum_{\beta \in \mathcal{A}: \beta \cap \alpha \neq \emptyset} \mu_\beta = \sum_{b=1}^k \sum_{\substack{\beta \cap \alpha \neq \emptyset \\ |\beta|=b}} \mu_\beta \leq \sum_{b=1}^k k \binom{n}{b-1} \mu_\beta.$$

Proof of Y2: (2.12): Suppose that $I_\alpha = 1$. Then H features no edge that contains a vertex in α and a vertex in β . If in addition $I_\beta^\alpha = 1$, then we obtain that $I_\beta = 1$ as well. Hence, $Y_\beta = Y_\beta^\alpha$.

(2.13): This just means that any two components of H are either disjoint or equal.

(2.14): To show that $Y_\gamma(Y_\beta - Y_\beta^\alpha) = 0$, assume that $I_\gamma = 1$. Then γ is a component of H , so that β cannot be a component, because $\gamma \neq \beta$ but $\gamma \cap \beta \neq \emptyset$; hence, $I_\beta = 0$. Furthermore, if γ is a component of H , then γ is also a component of $H \setminus \mathcal{E}(\alpha)$, so that $I_\gamma^\alpha = 1$. Consequently, $I_\beta^\alpha = 0$. Thus, $Y_\beta = Y_\beta^\alpha = 0$.

In order to prove that $Y_\gamma^\alpha(Y_\beta - Y_\beta^\alpha) = 0$, suppose that $I_\gamma^\alpha = 1$. Then $I_\beta^\alpha = 0$, because the intersecting sets β, γ cannot both be components of $H \setminus \mathcal{E}(\alpha)$. Therefore, we also have $I_\beta = 0$; for if β were a component of H , then β would also be a component of $H \setminus \mathcal{E}(\alpha)$. Hence, also in this case we obtain $Y_\beta = Y_\beta^\alpha = 0$.

Proof of Y3: Suppose that $I_\beta = 1$, i.e., β is a component of H . Then removing the edges $\mathcal{E}(\alpha)$ from H may cause β to split into several components B_1, \dots, B_l . Thus, if $Y_\gamma^\beta > 0$ for some $\gamma \in \mathcal{A}$ such that $\gamma \cap \beta \neq \emptyset$, then γ is one of the components B_1, \dots, B_l . Since $l \leq |\beta| \leq k$, this implies that given $I_\beta = 1$ we have the bound

$$\sum_{\gamma: \gamma \cap \beta \neq \emptyset, \gamma \cap \alpha = \emptyset} Y_\gamma^\alpha \leq k^2.$$

Hence, we obtain Y3.

Lemma 2.6. *Let $0 \leq l, r \leq 2$, and let $\alpha_1, \dots, \alpha_l, \beta_1, \dots, \beta_r \in \mathcal{A}$ be pairwise disjoint. Moreover, let $A_1, \dots, A_r, B_1, \dots, B_r \subset V$ be sets such that $A_i \subset B_i \subset V \setminus \beta_i$ and $|B_i| \leq 2k$ for all $1 \leq i \leq r$, and assume that $\bigcap_{i=1}^r B_i \setminus A_i = \emptyset$. Then*

$$\mathbb{P} \left[\bigwedge_{i=1}^l, \bigwedge_{j=1}^r I_{\alpha_i} = 1 \wedge I_{\beta_j}^{A_j} \neq I_{\beta_j}^{B_j} \right] \leq O(n^{-r} \cdot \text{polylog } n) \prod_{j=1}^l \mathbb{P}[I_{\alpha_i} = 1] \prod_{j=1}^r \mathbb{P}[I_{\beta_j} = 1].$$

This lemma easily implies **Y4–Y6**. Just note that the exponent of n occurring in **Y4–Y6** is equal to the number of factors of the form $Y_\beta - Y_\beta^\alpha$ while the number of μ_x occurring is equal to the total number of factors. Furthermore since $Y_\alpha \leq k$ we have $\mathbb{P}[I_\alpha = 1] = O(\mu_\alpha)$.

Proof. Since (2.29) entails that $I_{\beta_j}^{A_j} \neq I_{\beta_j}^{B_j} \leftrightarrow I_{\beta_j}^{B_j} = 1 \wedge I_{\beta_j}^{A_j} = 0$, we have

$$\mathbb{P} \left[\forall i, j : I_{\alpha_i} = 1 \wedge I_{\beta_j}^{A_j} \neq I_{\beta_j}^{B_j} \right] = \mathbb{P} \left[\forall i, j : I_{\alpha_i} = 1 \wedge I_{\beta_j}^{A_j} = 0 \wedge I_{\beta_j}^{B_j} = 1 \right]. \quad (2.30)$$

We shall bound the probability on the right hand side in terms of mutually independent events.

If $I_{\alpha_i} = 1$ and $I_{\beta_j}^{B_j} = 1$ for all i, j , then the hypergraphs induced on α_i and β_j are connected, i.e., the events $\mathcal{C}(\alpha_i)$ and $\mathcal{C}(\beta_j)$ occur. Note that these events are mutually independent, because $\mathcal{C}(\alpha_i)$ (resp. $\mathcal{C}(\beta_j)$) *only* depends on the presence of edges $e \in \mathcal{E}(\alpha_i) \setminus \mathcal{E}(V \setminus \alpha_i)$ (resp. $e \in \mathcal{E}(\beta_j) \setminus \mathcal{E}(V \setminus \beta_j)$).

Furthermore, if α_i is a component, then in H there occur no edges joining α_i and $V \setminus \alpha_i$; in other words, $H \cap \mathcal{E}(\alpha_i, V \setminus \alpha_i) = \emptyset$. However, these events are not necessarily independent, because $\mathcal{E}(\alpha_1, V \setminus \alpha_1)$ may contain edges that are incident with vertices in α_2 . Therefore, we consider the sets

$$\begin{aligned}\mathcal{F}(\alpha_i) &= \bigcup_{i' \neq i} \alpha_{i'} \cup \bigcup_{j=1}^r \beta_j \cup B_j, \mathcal{D}(\alpha_i) = \mathcal{E}(\alpha_i, V \setminus \alpha_i) \setminus \mathcal{E}(\mathcal{F}(\alpha_i)), \\ \mathcal{F}(\beta_j) &= \bigcup_{i=1}^l \alpha_i \cup \bigcup_{j' \neq j} \beta_{j'} \cup \bigcup_{j'=1}^r B_{j'}, \mathcal{D}(\beta_j) = \mathcal{E}(\beta_j, V \setminus \beta_j) \setminus \mathcal{E}(\mathcal{F}(\beta_j)).\end{aligned}$$

Then $I_{\alpha_i} = 1$ (resp. $I_{\beta_j}^{B_j} = 1$) implies that $\mathcal{D}(\alpha_i) \cap H = \emptyset$ (resp. $\mathcal{D}(\beta_j) \cap H = \emptyset$). Moreover, since the sets $\mathcal{D}(\alpha_i)$ and $\mathcal{D}(\beta_j)$ are pairwise disjoint, the events $\mathcal{D}(\alpha_i) \cap H = \emptyset$, $\mathcal{D}(\beta_j) \cap H = \emptyset$ are mutually independent.

Finally, we need to express the fact that $I_{\beta_j}^{A_j} = 0$ but $I_{\beta_j}^{B_j} = 1$. If this event occurs, then H contains an edge connecting β_j with $B_j \setminus A_j$, i.e., $H \cap \mathcal{E}(\beta_j, B_j \setminus A_j) \neq \emptyset$. Thus, let \mathcal{Q} denote the event that $H \cap \mathcal{E}(\beta_j, B_j \setminus A_j) \neq \emptyset$ for all $1 \leq j \leq r$.

Thus, we obtain

$$\begin{aligned}& \mathbb{P} \left[\forall i, j : I_{\alpha_i} = 1 \wedge I_{\beta_j}^{A_j} = 0 \wedge I_{\beta_j}^{B_j} = 1 \right] \\ & \leq \mathbb{P} \left[\bigwedge_{i=1}^l (\mathcal{C}(\alpha_i) \wedge (\mathcal{D}(\alpha_i) \cap H = \emptyset)) \wedge \bigwedge_{j=1}^r (\mathcal{C}(\beta_j) \wedge (\mathcal{D}(\beta_j) \cap H = \emptyset)) \wedge \mathcal{Q} \right] \\ & = \prod_{i=1}^l \mathbb{P}[\mathcal{C}(\alpha_i)] \mathbb{P}[\mathcal{D}(\alpha_i) \cap H = \emptyset] \times \prod_{j=1}^r \mathbb{P}[\mathcal{C}(\beta_j)] \mathbb{P}[\mathcal{D}(\beta_j) \cap H = \emptyset] \times \mathbb{P}[\mathcal{Q}].\end{aligned}\quad (2.31)$$

We shall prove below that

$$\mathbb{P}[\mathcal{D}(\alpha_i) \cap H = \emptyset] \sim (1-p)^{|\mathcal{E}(\alpha_i, V \setminus \alpha_i)|}, \quad \mathbb{P}[\mathcal{D}(\beta_j) \cap H = \emptyset] \sim (1-p)^{|\mathcal{E}(\beta_j, V \setminus \beta_j)|}, \quad (2.32)$$

$$\mathbb{P}[\mathcal{Q}] = O(n^{-r} \cdot \text{polylog } n). \quad (2.33)$$

Combining (2.28) and (2.30)–(2.33), we then obtain the assertion.

To establish (2.32), note that by definition $\mathcal{D}(\alpha_i) \subset \mathcal{E}(\alpha_i, V \setminus \alpha_i)$. Therefore,

$$\mathbb{P}[\mathcal{D}(\alpha_i) \cap H = \emptyset] = (1-p)^{|\mathcal{D}(\alpha_i)|} \geq (1-p)^{|\mathcal{E}(\alpha_i, V \setminus \alpha_i)|}. \quad (2.34)$$

On the other hand, we have $|\alpha_i|, |\mathcal{F}(\alpha_i)| = O(\text{polylog } n)$, and thus $|\mathcal{E}(\alpha_i, \mathcal{F}(\alpha_i))| \leq$

$|\alpha_i| \cdot |\mathcal{F}(\alpha_i)| \cdot \binom{n}{d-2} = O(n^{d-2} \cdot \text{polylog } n)$. Hence, as $p = O(n^{1-d})$, we obtain

$$\begin{aligned} \mathbb{P}[\mathcal{D}(\alpha_i) \cap H = \emptyset] &= (1-p)^{|\mathcal{D}(\alpha_i)|} \leq (1-p)^{|\mathcal{E}(\alpha_i, V \setminus \alpha_i)| - |\mathcal{E}(\alpha_i, \mathcal{F}(\alpha_i))|} \\ &\sim (1-p)^{|\mathcal{E}(\alpha_i, V \setminus \alpha_i)|} \exp(p \cdot O(n^{d-2} \cdot \text{polylog } n)) \\ &\sim (1-p)^{|\mathcal{E}(\alpha_i, V \setminus \alpha_i)|}. \end{aligned} \quad (2.35)$$

Combining (2.34) and (2.35), we conclude that $\mathbb{P}[\mathcal{D}(\alpha_i) \cap H = \emptyset] \sim (1-p)^{|\mathcal{E}(\alpha_i, V \setminus \alpha_i)|}$. As the same argument applies to $\mathbb{P}[\mathcal{D}(\beta_j) \cap H = \emptyset]$, we thus obtain (2.32).

Finally, we prove (2.33). If $r = 1$, then H contains an edge of $\mathcal{E}(\beta_1, B_1 \setminus A_1)$. Since

$$|\mathcal{E}(\beta_1, B_1 \setminus A_1)| \leq |\beta_1| \cdot |B_1 \setminus A_1| \cdot n^{d-2} = O(n^{d-2} \cdot \text{polylog } n),$$

and because each possible edge occurs with probability p independently, the probability of this event is $\mathbb{P}[\mathcal{Q}] \leq O(n^{d-2} \cdot \text{polylog } n) p = O(n^{-1} \cdot \text{polylog } n)$, as desired.

Now, assume that $r = 2$. Then H features edges $e_j \in \mathcal{E}(\beta_j, B_j \setminus A_j)$ ($j = 1, 2$).

1st case: $e_1 = e_2$. In this case, e_1 contains a vertex of each of the four sets $\beta_1, \beta_2, B_1 \setminus A_1, B_2 \setminus A_2$. Hence, the number of possible such edges is $\leq n^{d-4} \prod_{j=1}^2 |\beta_j| \cdot |B_j \setminus A_j| = O(n^{d-4} \cdot \text{polylog } n)$. Consequently, the probability that such an edge occurs in H is $\leq O(n^{d-4} \cdot \text{polylog } n) p = O(n^{-3} \cdot \text{polylog } n)$.

2nd case: $e_1 \neq e_2$. There are $\leq |\beta_j| \cdot |B_j \setminus A_j| \cdot n^{d-2} = O(n^{d-2} \cdot \text{polylog } n)$ ways to choose e_j ($j = 1, 2$). Hence, the probability that such edges e_1, e_2 occur in H is $\leq (O(n^{d-2} \cdot \text{polylog } n) p)^2 = O(n^{-2} \cdot \text{polylog } n)$.

Thus, in both cases we obtain the bound claimed in (2.33). \square

2.4 An Upper Bound for δ

In this section we show that the conditions **Y1–Y6** provided in Lemma 2.4 suffice to prove $\delta = O(\sigma^{-1})$, thereby proving Lemma 2.4

Lemma 2.7. $\sum_{\alpha \in \mathcal{A}} \mathbb{E}[|X_\alpha| Z_\alpha^2] = O(\sigma^{-3} \mathbb{E}[Y] \cdot \text{polylog } n)$

Proof. Let

$$\begin{aligned} S_1 &= \sum_{\alpha \in \mathcal{A}} \mathbb{E} \left[Y_\alpha \left(\sum_{\alpha \cap \beta \neq \emptyset} Y_\beta \right)^2 \right], S_2 = \sum_{\alpha \in \mathcal{A}} \mathbb{E} \left[\mu_\alpha \left(\sum_{\alpha \cap \beta \neq \emptyset} Y_\beta \right)^2 \right], \\ S_3 &= \sum_{\alpha \in \mathcal{A}} \mathbb{E} \left[Y_\alpha \left(\sum_{\alpha \cap \beta = \emptyset} (Y_\beta - Y_\beta^\alpha) \right)^2 \right], S_4 = \sum_{\alpha \in \mathcal{A}} \mathbb{E} \left[\mu_\alpha \left(\sum_{\alpha \cap \beta = \emptyset} (Y_\beta - Y_\beta^\alpha) \right)^2 \right]. \end{aligned}$$

Since $X_\alpha = (Y_\alpha - \mu_\alpha)/\sigma \leq (Y_\alpha + \mu_\alpha)/\sigma$, (2.8) entails that

$$\begin{aligned} \mathbb{E} \left[|X_\alpha|^2 \right] &\leq 2\sigma^{-3} \mathbb{E} \left[(Y_\alpha + \mu_\alpha) \left(\left(\sum_{\alpha \cap \beta \neq \emptyset} Y_\beta \right)^2 + \left(\sum_{\alpha \cap \beta = \emptyset} (Y_\beta - Y_\beta^\alpha) \right)^2 \right) \right] \\ &\leq 2\sigma^{-3} (S_1 + S_2 + S_3 + S_4). \end{aligned}$$

Therefore, it suffices to show that $S_j = O(\mathbb{E}[Y] \cdot \text{polylog } n)$ for $j = 1, 2, 3, 4$.

Regarding S_1 , we obtain the bound

$$S_1 = \sum_{\alpha \in \mathcal{A}} \sum_{\alpha \cap \beta \neq \emptyset} \sum_{\alpha \cap \gamma \neq \emptyset} \mathbb{E}[Y_\alpha Y_\beta Y_\gamma] \stackrel{(2.13)}{\leq} k^2 \sum_{\alpha \in \mathcal{A}} \mathbb{E}[Y_\alpha] \leq O(\mathbb{E}[Y] \cdot \text{polylog } n).$$

With respect to S_2 , note that due to (2.13) and (2.15) we have $\mathbb{E}[Y_\beta Y_\gamma] \leq k\mu_\beta$ if $\beta = \gamma$, $\mathbb{E}[Y_\beta Y_\gamma] = 0$ if $\beta \neq \gamma$ but $\beta \cap \gamma \neq \emptyset$, and $\mathbb{E}[Y_\beta Y_\gamma] = O(\mu_\beta \mu_\gamma \cdot \text{polylog } n)$ if $\beta \cap \gamma = \emptyset$. Consequently,

$$\begin{aligned} S_2 &= \sum_{\alpha \in \mathcal{A}} \mu_\alpha \sum_{\alpha \cap \beta \neq \emptyset} \sum_{\alpha \cap \gamma \neq \emptyset} \mathbb{E}[Y_\beta Y_\gamma] \\ &\leq \sum_{\alpha \in \mathcal{A}} \mu_\alpha \sum_{\alpha \cap \beta \neq \emptyset} \sum_{\alpha \cap \gamma \neq \emptyset} O(\mu_\beta \mu_\gamma \cdot \text{polylog } n) \stackrel{\mathbf{Y1}}{\leq} O(\mathbb{E}[Y] \cdot \text{polylog } n). \end{aligned} \quad (2.36)$$

Concerning S_3 , we obtain

$$\begin{aligned} S_3 &= \sum_{\alpha \in \mathcal{A}} \sum_{\alpha \cap \beta = \emptyset} \sum_{\alpha \cap \gamma = \emptyset} \mathbb{E} \left[Y_\alpha (Y_\beta - Y_\beta^\alpha) (Y_\gamma - Y_\gamma^\alpha) \right] \\ &\stackrel{(2.21), (2.14)}{\leq} \sum_{\alpha \in \mathcal{A}} \sum_{\alpha \cap \beta = \emptyset} \sum_{\alpha \cap \gamma = \emptyset} O(\mu_\alpha \mu_\beta \mu_\gamma n^{-2} \cdot \text{polylog } n) \\ &\leq O(n^{-2} \cdot \text{polylog } n) \mathbb{E}[Y]^3 \stackrel{\mathbf{Y1}}{\leq} O(\mathbb{E}[Y] \cdot \text{polylog } n). \end{aligned}$$

To bound S_4 , we note that for disjoint $\alpha, \beta \in \mathcal{A}$ and $\gamma \in \mathcal{A}$ disjoint from α the conditions (2.16), (2.13), and (2.22) yield

$$\mathbb{E} \left[|(Y_\beta - Y_\beta^\alpha)(Y_\gamma - Y_\gamma^\alpha)| \right] = \begin{cases} O\left(\frac{\mu_\beta}{n} \cdot \text{polylog } n\right) & \text{if } \beta = \gamma \\ 0 & \text{if } \beta \neq \gamma, \beta \cap \gamma \neq \emptyset \\ O\left(\frac{\mu_\beta \mu_\gamma}{n^2} \cdot \text{polylog } n\right) & \text{if } \beta \cap \gamma = \emptyset. \end{cases}$$

Therefore,

$$\begin{aligned} &\sum_{\alpha \cap \beta = \emptyset} \sum_{\alpha \cap \gamma = \emptyset} \mathbb{E} \left[|(Y_\beta - Y_\beta^\alpha)(Y_\gamma - Y_\gamma^\alpha)| \right] \\ &\leq \sum_{\beta \in \mathcal{A}} \sum_{\gamma \in \mathcal{A}} O\left(\frac{\mu_\beta \mu_\gamma}{n^2} \cdot \text{polylog } n\right) + \sum_{\beta \in \mathcal{A}} O\left(\frac{\mu_\beta}{n} \cdot \text{polylog } n\right) \\ &\leq O(\mathbb{E}[Y]^2/n^2 \cdot \text{polylog } n) + O(\mathbb{E}[Y]/n \cdot \text{polylog } n) \\ &= O(\text{polylog } n). \end{aligned}$$

Hence, we obtain

$$S_4 \leq \sum_{\alpha \in \mathcal{A}} \mu_\alpha \sum_{\alpha \cap \beta \neq \emptyset} \sum_{\alpha \cap \gamma \neq \emptyset} \mathbb{E} \left[(Y_\beta - Y_\beta^\alpha)(Y_\gamma - Y_\gamma^\alpha) \right] \leq O(\mathbb{E}[Y] \cdot \text{polylog } n)$$

□

Lemma 2.8. $\sum_{\alpha \in \mathcal{A}} \sum_{\beta \in \mathcal{A}} \mathbb{E} [|X_\alpha Z_{\alpha\beta} V_{\alpha\beta}|] = O(\sigma^{-3} \mathbb{E}[Y] \cdot \text{polylog } n)$

Proof. Let $S_1 = \sum_{\alpha \cap \beta \neq \emptyset} \mathbb{E} [|X_\alpha Y_\beta V_{\alpha\beta}|]$ and $S_2 = \sum_{\alpha \cap \beta = \emptyset} \mathbb{E} [|X_\alpha (Y_\beta - Y_\beta^\alpha) V_{\alpha\beta}|]$. Then the definition (2.9) of $Z_{\alpha\beta}$ yields that $\sum_{\alpha \in \mathcal{A}} \sum_{\beta \in \mathcal{A}} \mathbb{E} [|X_\alpha Z_{\alpha\beta} V_{\alpha\beta}|] \leq \sigma^{-1}(S_1 + S_2)$. Hence, it suffices to show that $S_1, S_2 = O(\sigma^{-2} \mathbb{E}[Y] \cdot \text{polylog } n)$.

To bound S_1 , we note that $Y_\alpha Y_\beta = 0$ if $\alpha \cap \beta \neq \emptyset$ but $\alpha \neq \beta$ by (2.13), and that $V_{\alpha\beta} = 0$ if $\alpha = \beta$ by the definition (2.10) of $V_{\alpha\beta}$. Thus, if $\alpha \cap \beta \neq \emptyset$, then

$$\mathbb{E} [|X_\alpha Y_\beta V_{\alpha\beta}|] \stackrel{(2.8)}{\leq} \sigma^{-1} \mathbb{E} [(Y_\alpha + \mu_\alpha) |Y_\beta V_{\alpha\beta}|] \leq \sigma^{-1} \mu_\alpha \mathbb{E} [|Y_\beta V_{\alpha\beta}|]. \quad (2.37)$$

Furthermore,

$$T_1(\alpha, \beta) = \sum_{\gamma: \gamma \cap \beta \neq \emptyset, \gamma \cap \alpha = \emptyset} \mathbb{E} [|Y_\beta Y_\gamma^\alpha|] \stackrel{\text{Y7}}{\leq} k^2 \mu_\beta. \quad (2.38)$$

$$T_2(\alpha) = \sum_{\alpha \cap \beta \neq \emptyset} \sum_{\substack{\gamma: \beta \cap \gamma = \emptyset \\ \wedge \alpha \cap \gamma = \emptyset}} \mathbb{E} [|Y_\beta |Y_\gamma^\alpha - Y_\gamma^{\alpha \cup \beta}|] \quad (2.39)$$

$$\stackrel{(2.18)}{\leq} \sum_{\beta: \alpha \cap \beta \neq \emptyset} \sum_{\substack{\gamma: \beta \cap \gamma = \emptyset \\ \wedge \alpha \cap \gamma = \emptyset}} O\left(\frac{\mu_\beta \mu_\gamma}{n} \cdot \text{polylog } n\right)$$

$$\leq O(n^{-1} \cdot \text{polylog } n) \left(\sum_{\gamma \in \mathcal{A}} \mu_\gamma \right) \sum_{\alpha \cap \beta \neq \emptyset} \mu_\beta$$

$$\stackrel{\text{Y1}}{\leq} O(n^{-1} \mathbb{E}[Y] \cdot \text{polylog } n) = O(\text{polylog } n). \quad (2.40)$$

Combining (2.37)–(2.40), we get

$$\begin{aligned} S_1 &\leq \sigma^{-1} \sum_{\alpha \in \mathcal{A}} \sum_{\alpha \cap \beta \neq \emptyset} \mu_\alpha \mathbb{E} [|Y_\beta V_{\alpha\beta}|] \stackrel{(2.10)}{\leq} \sigma^{-2} \sum_{\alpha \in \mathcal{A}} \mu_\alpha \left(T_2(\alpha) + \sum_{\beta: \alpha \cap \beta \neq \emptyset} T_1(\alpha, \beta) \right) \\ &\leq O(\sigma^{-2} \cdot \text{polylog } n) \left(\mathbb{E}[Y] + k^2 \sum_{\beta: \alpha \cap \beta \neq \emptyset} \mu_\beta \right) \stackrel{\text{Y1}}{\leq} O(\sigma^{-2} \mathbb{E}[Y] \cdot \text{polylog } n) \end{aligned}$$

To bound S_2 , let $\alpha, \beta \in \mathcal{A}$ be disjoint. As $X_\alpha \leq (Y_\alpha + \mu_\alpha)/\sigma$, we obtain

$$\begin{aligned}
 \mathbb{E} \left[\left| X_\alpha (Y_\beta - Y_\beta^\alpha) V_{\alpha\beta} \right| \right] &\leq \sigma^{-1} \mathbb{E} \left[\left| (Y_\alpha + \mu_\alpha) (Y_\beta - Y_\beta^\alpha) V_{\alpha\beta} \right| \right] \\
 &\stackrel{(2.10), (2.14)}{\leq} \sigma^{-2} \mathbb{E} \left[\left| (Y_\alpha + \mu_\alpha) (Y_\beta - Y_\beta^\alpha) Y_\beta^\alpha \right| \right] \\
 &\quad + \sigma^{-2} \sum_{\substack{\gamma: \beta \cap \gamma = \emptyset \\ \wedge \alpha \cap \gamma = \emptyset}} \mathbb{E} \left[\left| (Y_\alpha + \mu_\alpha) (Y_\beta - Y_\beta^\alpha) (Y_\gamma^\alpha - Y_\gamma^{\alpha \cup \beta}) \right| \right] \\
 &\leq \sigma^{-2} (T_1 + T_2 + T_3 + T_4), \tag{2.41}
 \end{aligned}$$

where

$$\begin{aligned}
 T_1 &= \mathbb{E} \left[\left| Y_\alpha (Y_\beta - Y_\beta^\alpha) Y_\beta^\alpha \right| \right], \\
 T_2 &= \mu_\alpha \mathbb{E} \left[\left| (Y_\beta - Y_\beta^\alpha) Y_\beta^\alpha \right| \right], \\
 T_3 &= \sum_{\substack{\gamma: \beta \cap \gamma = \emptyset \\ \wedge \alpha \cap \gamma = \emptyset}} \mathbb{E} \left[\left| Y_\alpha (Y_\beta - Y_\beta^\alpha) (Y_\gamma^\alpha - Y_\gamma^{\alpha \cup \beta}) \right| \right], \\
 T_4 &= \mu_\alpha \sum_{\substack{\gamma: \beta \cap \gamma = \emptyset \\ \wedge \alpha \cap \gamma = \emptyset}} \mathbb{E} \left[\left| (Y_\beta - Y_\beta^\alpha) (Y_\gamma^\alpha - Y_\gamma^{\alpha \cup \beta}) \right| \right].
 \end{aligned}$$

Now, $T_1 = 0$ by (2.12). Moreover, bounding T_2 by (2.16), T_3 by (2.20) and T_4 by (2.19), we obtain

$$\begin{aligned}
 \sigma^2 \mathbb{E} \left[\left| X_\alpha (Y_\beta - Y_\beta^\alpha) V_{\alpha\beta} \right| \right] &\leq O \left(\frac{\mu_\alpha \mu_\beta}{n} \cdot \text{polylog } n \right) + \sum_{\substack{\gamma: \beta \cap \gamma = \emptyset \\ \wedge \alpha \cap \gamma = \emptyset}} O \left(\frac{\mu_\alpha \mu_\beta \mu_\gamma}{n^2} \cdot \text{polylog } n \right) \\
 &= O \left(\frac{\mu_\alpha \mu_\beta}{n} \cdot \text{polylog } n \right).
 \end{aligned}$$

Thus, (2.41) yields

$$\begin{aligned}
 S_2 &\leq \sigma^{-2} \sum_{\alpha \cap \beta = \emptyset} O \left(\frac{\mu_\alpha \mu_\beta}{n} \cdot \text{polylog } n \right) = O \left(n^{-1} \sigma^{-2} \mathbb{E} [Y]^2 \cdot \text{polylog } n \right) \\
 &= O \left(\sigma^{-2} \mathbb{E} [Y] \cdot \text{polylog } n \right),
 \end{aligned}$$

as desired. \square

Lemma 2.9. $\sum_{\alpha \in \mathcal{A}} \sum_{\beta \in \mathcal{A}} \mathbb{E} [|X_\alpha Z_{\alpha\beta}|] \mathbb{E} [|Z_\alpha + V_{\alpha\beta}|] = O(\sigma^{-3} \mathbb{E} [Y] \cdot \text{polylog } n)$

Proof. Since $|\sigma X_\alpha| \leq Y_\alpha + \mu_\alpha$,

$$\begin{aligned}
 \sum_{\alpha \in \mathcal{A}} \sum_{\beta \in \mathcal{A}} \mathbb{E} [|X_\alpha Z_{\alpha\beta}|] \mathbb{E} [|Z_\alpha + V_{\alpha\beta}|] &\leq \sigma^{-1} \left(\sum_{\alpha \in \mathcal{A}} \sum_{\beta \in \mathcal{A}} \mu_\alpha \mathbb{E} [|Z_{\alpha\beta}|] (\mathbb{E} [|Z_\alpha|] + \mathbb{E} [|V_{\alpha\beta}|]) + \right. \\
 &\quad \left. \mathbb{E} [Y_\alpha | Z_{\alpha\beta}|] (\mathbb{E} [|Z_\alpha|] + \mathbb{E} [|V_{\alpha\beta}|]) \right). \tag{2.42}
 \end{aligned}$$

Furthermore, we have the three estimates

$$\begin{aligned} \sigma \mathbb{E} [|Z_\alpha|] &\leq \sigma \sum_{\beta \in \mathcal{A}} \mathbb{E} [|Z_{\alpha\beta}|] \stackrel{(2.9)}{=} \sum_{\alpha \cap \beta \neq \emptyset} \mu_\beta + \sum_{\alpha \cap \beta = \emptyset} \mathbb{E} [|Y_\beta - Y_\beta^\alpha|] \\ &\stackrel{(2.16), \mathbf{Y1}}{\leq} \sum_{\beta \in \mathcal{A}} O(n^{-1} \mu_\beta \cdot \text{polylog } n) = O(\text{polylog } n), \end{aligned} \quad (2.43)$$

$$\begin{aligned} \sigma \mathbb{E} [|V_{\alpha\beta}|] &\stackrel{(2.10)}{\leq} \sum_{\substack{\gamma: \beta \cap \gamma \neq \emptyset \\ \wedge \alpha \cap \gamma = \emptyset}} \mathbb{E} [|Y_\gamma^\alpha|] + \sum_{\substack{\gamma: \beta \cap \gamma = \emptyset \\ \wedge \alpha \cap \gamma = \emptyset}} \mathbb{E} [|Y_\gamma^\alpha - Y_\gamma^{\alpha \cup \beta}|] \\ &\stackrel{(2.23), \mathbf{Y1}}{=} \sum_{\gamma \in \mathcal{A}} O(n^{-1} \mu_\gamma \cdot \text{polylog } n) \leq O(\text{polylog } n), \end{aligned} \quad (2.44)$$

$$\begin{aligned} \sum_{\beta \in \mathcal{A}} \sigma \mathbb{E} [Y_\alpha | Z_{\alpha\beta}] &\stackrel{(2.9)}{=} \sum_{\alpha \cap \beta \neq \emptyset} \mathbb{E} [Y_\alpha Y_\beta] + \sum_{\alpha \cap \beta = \emptyset} \mathbb{E} [Y_\alpha | Y_\beta - Y_\beta^\alpha] \\ &\stackrel{(2.13), (2.17)}{=} k\mu_\alpha + \sum_{\alpha \cap \beta = \emptyset} \frac{\mu_\alpha \mu_\beta}{n} = O(\mu_\alpha \cdot \text{polylog } n). \end{aligned} \quad (2.45)$$

Now, (2.43)–(2.45) yield

$$\begin{aligned} \sum_{\alpha \in \mathcal{A}} \sum_{\beta \in \mathcal{A}} \mu_\alpha \mathbb{E} [|Z_{\alpha\beta}|] (\mathbb{E} [|Z_\alpha|] + \mathbb{E} [|V_{\alpha\beta}|]) &= O(\sigma^{-2} \cdot \text{polylog } n) \sum_{\alpha \in \mathcal{A}} \mu_\alpha \\ &= O(\sigma^{-2} \mathbb{E} [Y] \cdot \text{polylog } n), \end{aligned} \quad (2.46)$$

$$\begin{aligned} \sum_{\alpha \in \mathcal{A}} \sum_{\beta \in \mathcal{A}} \mathbb{E} [Y_\alpha | Z_{\alpha\beta}] (\mathbb{E} [|Z_\alpha|] + \mathbb{E} [|V_{\alpha\beta}|]) &= O(\sigma^{-2} \cdot \text{polylog } n) \sum_{\alpha \in \mathcal{A}} \mu_\alpha \\ &= O(\sigma^{-2} \mathbb{E} [Y] \cdot \text{polylog } n). \end{aligned} \quad (2.47)$$

Combining (2.42), (2.46), and (2.47), we obtain the assertion. \square

Finally, Lemma 2.4 is an immediate consequence of Lemmas 2.7–2.9.

Chapter 3

A Local Limit Theorem for the Number of Vertices

3.1 Results

Though Theorem 2.1 provides useful information about the distribution of $\mathcal{N}(H_d(n, p))$ and may be sufficiently accurate in many contexts, the main result of this thesis is actually a *local limit theorem* for $\mathcal{N}(H_d(n, p))$, which characterises the distribution of $\mathcal{N}(H_d(n, p))$ even more precisely. To motivate the local limit theorem, we emphasise that Theorem 2.1 only estimates $\mathcal{N}(G_{n,p})$ up to an error of $o(\sigma)$, where $\sigma = \Theta(\sqrt{n})$. That is, we do obtain from (2.2) that for arbitrarily small but fixed $\delta > 0$

$$\mathbb{P}[|\mathcal{N}(H_d(n, p)) - \nu| \leq \delta\sigma] \sim \frac{1}{\sqrt{2\pi}\sigma} \int_{-\delta\sigma}^{\delta\sigma} \exp\left(-\frac{(\nu - (1-\rho)n - t)^2}{2\sigma^2}\right) dt, \quad (3.1)$$

i.e., we can estimate the probability that $\mathcal{N}(H_d(n, p))$ deviates from some value ν by at most $\delta\sigma$. However, it is impossible to derive from (2.2) or (3.1) the asymptotic probability that $\mathcal{N}(H_d(n, p))$ *hits* ν *exactly*.

By contrast, our next theorem shows that for any integer ν such that $|\nu - (1-\rho)n| \leq O(\sigma)$ we have

$$\mathbb{P}[\mathcal{N}(H_d(n, p)) = \nu] \sim \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(\nu - (1-\rho)n)^2}{2\sigma^2}\right), \quad (3.2)$$

provided that $(d-1)^{-1} + \varepsilon \leq \binom{n-1}{d-1}p = O(1)$. Note that (3.2) is exactly what we would obtain from (3.1) if we were allowed to set $\delta = \frac{1}{2}\sigma(n, p)^{-1}$ in that equation. Stated rigorously, the local limit theorem reads as follows.

Theorem 3.1. *Let $d \geq 2$ be a fixed integer. For any two compact intervals $\mathcal{I} \subset \mathbb{R}$, $\mathcal{J} \subset ((d-1)^{-1}, \infty)$, and for any $\delta > 0$ there exists $n_0 > 0$ such that the following holds. Let $p = p(n)$ be a sequence such that $c = c(n) = \binom{n-1}{d-1}p \in \mathcal{J}$ for all n , then $\mathbb{P}[\mathcal{N}(H_d(n, p)) = \nu] = O(n^{-1/2})$ for all ν .*

Furthermore let $p = p(n)$ be a sequence such that $c = c(n) = \binom{n-1}{d-1}p \in \mathcal{J}$ for all n , let $0 < \rho = \rho(n) < 1$ be the unique solution to (1.2), and let σ be as in (2.1). If $n \geq n_0$ and if ν is an integer such that $\sigma^{-1}(\nu - (1 - \rho)n) \in \mathcal{I}$, then

$$\frac{1 - \delta}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(\nu - (1 - \rho)n)^2}{2\sigma^2}\right) \leq \mathbb{P}[\mathcal{N}(H_d(n, p)) = \nu] \leq \frac{1 + \delta}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(\nu - (1 - \rho)n)^2}{2\sigma^2}\right).$$

3.2 Proof of the Local Limit Theorem

Throughout this section, we assume that $c = c(n) = \binom{n-1}{d-1}p \in \mathcal{J}$ for some compact interval $\mathcal{J} \subset ((d-1)^{-1}, \infty)$. Moreover, we let $\mathcal{I} \subset \mathbb{R}$ be some fixed compact interval, and ν denotes an integer such that $(\nu - (1 - \rho)n)/\sigma \in \mathcal{I}$. All asymptotics are understood to hold uniformly in c and $(\nu - (1 - \rho)n)/\sigma$.

3.2.1 Outline

Let $\varepsilon = \varepsilon(\mathcal{J}) > 0$ be independent of n and small enough so that $(1 - \varepsilon)\binom{n-1}{d-1}p > (d-1)^{-1} + \varepsilon$. Set $p_1 = (1 - \varepsilon)p$. Moreover, let p_2 be the solution to the equation $p_1 + p_2 - p_1p_2 = p$; then $p_2 \sim \varepsilon p$. We expose the edges of $H_d(n, p)$ in four “rounds” as follows.

- R1.** As a first step, we let H_1 be a random hypergraph obtained by including each of the $\binom{n}{d}$ possible edges with probability p_1 independently. Let G denote the largest component of H_1 .
- R2.** Let H_2 be the hypergraph obtained from H_1 by adding each edge $e \notin H_1$ that lies completely outside of G (i.e., $e \subset V \setminus G$) with probability p_2 independently.
- R3.** Obtain H_3 by adding each possible edge $e \notin H_1$ that contains vertices of both G and $V \setminus G$ with probability p_2 independently.
- R4.** Finally, include each possible edge $e \notin H_1$ such that $e \subset G$ with probability p_2 independently.

Here the 1st round corresponds to the first portion of edges mentioned in Section 1.2, and the edges added in the 2nd–4th round correspond to the second portion. Note that for each possible edge $e \subset V$ the probability that e is actually present in H_4 is $p_1 + (1 - p_1)p_2 = p$, hence $H_4 = H_d(n, p)$. Moreover, as $\binom{n-1}{d-1}p_1 > (d-1)^{-1} + \varepsilon$ by our choice of ε , Theorem 1.2 entails that a.a.s. H_1 has exactly one largest component of linear order $\Omega(n)$ (the giant). Further, the edges added in the 4th round do not affect the order of the largest component, i.e., $\mathcal{N}(H_4) = \mathcal{N}(H_3)$.

In order to analyse the distribution of $\mathcal{N}(H_d(n, p))$, we first establish *central limit theorems* for $\mathcal{N}(H_1) = |G|$ and $\mathcal{N}(H_3) = \mathcal{N}(H_4) = \mathcal{N}(H_d(n, p))$, i.e., we prove that (centralised and normalised versions of) $\mathcal{N}(H_1)$ and $\mathcal{N}(H_3)$ are asymptotically normal. Then, we investigate the number of vertices $\mathcal{S} = \mathcal{N}(H_3) - \mathcal{N}(H_1)$ that get attached to G

during the 3rd round. We shall prove that *given that* $|G| = n_1$, \mathcal{S} is *locally* normal with mean $\mu_{\mathcal{S}} + (n_1 - \mu_1)\lambda_{\mathcal{S}}$ and variance $\sigma_{\mathcal{S}}^2$ independent of n_1 . Finally, we combine these results to obtain the local limit theorem for $\mathcal{N}(H_d(n, p)) = \mathcal{N}(H_3) = \mathcal{N}(H_1) + \mathcal{S}$.

Let $c_1 = \binom{n-1}{d-1}p_1$ and $c_3 = \binom{n-1}{d-1}p$. Moreover, let $0 < \rho_3 < \rho_1 < 1$ signify the solutions to the transcendental equations $\rho_j = \exp\left(c_j(\rho_j^{d-1} - 1)\right)$ and set for $j = 1, 3$

$$\mu_j = (1 - \rho_j)n, \quad \sigma_j^2 = \frac{\rho_j(1 - \rho_j + c_j(d-1)(\rho_j - \rho_j^{d-1}))n}{(1 - c_j(d-1)\rho_j^{d-1})^2} \quad (\text{cf. Theorem 1.2}).$$

The following proposition, which is a corollary to Theorem 2.1, establishes a central limit theorem for both $\mathcal{N}(H_1)$ and $\mathcal{N}(H_3)$.

Proposition 3.2. *$(\mathcal{N}(H_j) - \mu_j)/\sigma_j$ converges in distribution to the standard normal distribution for $j = 1, 3$.*

Let now and in the following $\mathcal{N}_1 = \mathcal{N}(H_1)$ and $\mathcal{N}_3 = \mathcal{N}(H_3)$. With respect to the distribution of \mathcal{S} , we will establish the following local limit theorem in Section 3.3.

Proposition 3.3. *Suppose that $|n_1 - \mu_1| \leq n^{0.6}$.*

1. *The conditional expectation of \mathcal{S} given that $|G| = n_1$ satisfies $\mathbb{E}[\mathcal{S}|\mathcal{N}_1 = n_1] = \mu_{\mathcal{S}} + \lambda_{\mathcal{S}}(n_1 - \mu_1) + o(\sqrt{n})$, where $\mu_{\mathcal{S}} = \Theta(n)$ and $\lambda_{\mathcal{S}} = \Theta(1)$ are independent of n_1 .*
2. *There is a constant $C > 0$ such that for all s satisfying $|\mu_{\mathcal{S}} + \lambda_{\mathcal{S}}(n_1 - \mu_1) - s| \leq n^{0.6}$ we have $\mathbb{P}[\mathcal{S} = \nu|\mathcal{N}_1 = n_1] \leq Cn^{-\frac{1}{2}}$.*
3. *If s is an integer such that $|\mu_{\mathcal{S}} + \lambda_{\mathcal{S}}(n_1 - \mu_1) - s| \leq O(\sqrt{n})$, then*

$$\mathbb{P}[\mathcal{S} = s|\mathcal{N}_1 = n_1] \sim \frac{1}{\sqrt{2\pi}\sigma_{\mathcal{S}}} \exp\left(-\frac{(\mu_{\mathcal{S}} + \lambda_{\mathcal{S}}(n_1 - \mu_1) - s)^2}{2\sigma_{\mathcal{S}}^2}\right),$$

where $\sigma_{\mathcal{S}} = \Theta(\sqrt{n})$ is independent of n_1 .

Since $\mathcal{N}_3 = \mathcal{N}_1 + \mathcal{S}$, Propositions 3.2 and 3.3 yield

$$\mu_3 = \mu_1 + \mu_{\mathcal{S}} + o(\sqrt{n}). \quad (3.3)$$

Combining Propositions 3.2 and 3.3, we derive the following formula for $\mathbb{P}[\mathcal{N}_3 = \nu]$ in Section 3.2.2. Recall that we are assuming that ν is an integer such that $(\nu - \mu)/\sigma = (\nu - \mu_3)/\sigma_3 \in \mathcal{I}$.

Corollary 3.4. *Letting $z = (\nu - \mu_3)/\sigma_3$, we have*

$$\mathbb{P}[\mathcal{N}_3 = \nu] \sim \frac{1}{2\pi\sigma_{\mathcal{S}}} \int_{-\infty}^{\infty} \exp\left(-\frac{x^2}{2} - \frac{1}{2}\left((x \cdot (1 + \lambda_{\mathcal{S}})\frac{\sigma_1}{\sigma_{\mathcal{S}}} - z \cdot \frac{\sigma_3}{\sigma_{\mathcal{S}}})^2\right)\right) dx. \quad (3.4)$$

Proof of Theorem 3.1. Integrating the right hand side of (3.4), we obtain an expression of the form

$$\mathbb{P}[\mathcal{N}_3 = \nu] \sim \frac{1}{\sqrt{2\pi\tau}} \exp\left(-\frac{(\nu - \kappa)^2}{2\tau^2}\right), \quad (3.5)$$

where $\kappa, \tau^2 = \Theta(n)$. Therefore, on the one hand $(\mathcal{N}_3 - \mu_3)/\sigma_3$ converges in distribution to the normal distribution with mean $\kappa - \mu_3$ and variance $(\tau/\sigma_3)^2$. On the other hand, Proposition 3.2 states that $(\mathcal{N}_3 - \mu_3)/\sigma_3$ converges to the standard normal distribution. Consequently, $|\kappa - \mu_3| = o(\tau)$ and $\tau \sim \sigma_3$. Plugging these estimates into (3.5), we obtain $\mathbb{P}[\mathcal{N}_3 = \nu] \sim \frac{1}{\sqrt{2\pi\sigma_3}} \exp\left(-\frac{1}{2}(\nu - \mu_3)^2\sigma_3^{-2}\right)$. Since $\mathcal{N}_3 = \mathcal{N}(H_d(n, p))$, this yields the assertion. \square

3.2.2 The Distribution of \mathcal{N}_3 as a Combination of \mathcal{N}_1 and \mathcal{S}

Proof of Corollary 3.4. Let $\alpha > 0$ be arbitrarily small but fixed as $n \rightarrow \infty$, and let $C' = C'(\alpha) > 0$ be a large enough number depending only on α . Set $J = \{n_1 \in \mathbb{Z} : |n_1 - \mu_1| \leq C'\sqrt{n}\}$, let $J' = \{n_1 \in \mathbb{Z} : C'\sqrt{n} < |n_1 - \mu_1| \leq n^{0.6}\}$, and $J'' = \{n_1 \in \mathbb{Z} : |n_1 - \mu_1| > n^{0.6}\}$. Then letting

$$\Psi_X = \sum_{n_1 \in X} \mathbb{P}[\mathcal{N}_1 = n_1] \mathbb{P}[\mathcal{S} = \nu - n_1 | \mathcal{N}_1 = n_1], \quad \text{for } X \in \{J, J', J''\}$$

we have $\mathbb{P}[\mathcal{N}_3 = \nu] = \Psi_J + \Psi_{J'} + \Psi_{J''}$, and we shall estimate each of the three summands individually.

Since Theorem 1.2 implies that $\mathbb{P}[|\mathcal{N}_1 - \mu_1| > n^{0.51}] \leq n^{-100}$, we conclude that

$$\Psi_{J''} \leq \mathbb{P}[\mathcal{N}_1 \in J''] \leq n^{-100}. \quad (3.6)$$

Furthermore, as $\sigma_1^2 = O(n)$, Chebyshev's inequality implies that

$$\mathbb{P}[\mathcal{N}_1 \in J'] \leq \mathbb{P}[|\mathcal{N}_1 - \mu_1| > C'\sqrt{n}] \leq \sigma_1^2 C'^{-2} n^{-1} < \alpha/C', \quad (3.7)$$

provided that C' is large enough. Hence, combining (3.7) with the second part of Proposition 3.3, we obtain

$$\Psi_{J'} \leq \mathbb{P}[\mathcal{N}_1 \in J'] \cdot \frac{C}{\sqrt{n}} \leq \frac{\alpha C}{C' \sqrt{n}} < \alpha n^{-1/2}, \quad (3.8)$$

where we need to pick C' sufficiently large.

To estimate the contribution of $n_1 \in J$, we split J into subintervals J_1, \dots, J_K of length between $\frac{\sigma_1}{2C'}$ and $\frac{\sigma_1}{C'}$. Moreover, let I_j be the interval $[(\min J_j - \mu_1)/\sigma_1, (\max J_j - \mu_1)/\sigma_1]$. Then Proposition 3.2 implies that

$$\frac{1 - \alpha}{\sqrt{2\pi}} \int_{I_j} \exp(-x^2/2) dx \leq \sum_{n_1 \in J_j} \mathbb{P}[\mathcal{N}_1 = n_1] \leq \frac{1 + \alpha}{\sqrt{2\pi}} \int_{I_j} \exp(-x^2/2) dx \quad (3.9)$$

for each $1 \leq j \leq K$. Furthermore, Proposition 3.3 yields

$$\mathbb{P}[\mathcal{S} = \nu - n_1 | \mathcal{N}_1 = n_1] \sim \frac{1}{\sqrt{2\pi}\sigma_{\mathcal{S}}} \exp\left(-\frac{(\nu - n_1 - \mu_{\mathcal{S}} - \lambda_{\mathcal{S}}(n_1 - \mu_1))^2}{2\sigma_{\mathcal{S}}^2}\right).$$

for each $n_1 \in J$. Hence, choosing C' sufficiently large, we can achieve that for all $n_1 \in J_j$ and all $x \in I_j$ the bound

$$\begin{aligned} \mathbb{P}[\mathcal{S} = \nu - n_1 | \mathcal{N}_1 = n_1] &\leq \frac{(1+\alpha)^2}{\sqrt{2\pi}\sigma_{\mathcal{S}}} \exp\left(-\frac{(\nu - \mu_1 - \sigma_1 x - \mu_{\mathcal{S}} - \lambda_{\mathcal{S}}(n_1 - \mu_1))^2}{2\sigma_{\mathcal{S}}^2}\right) \\ &\stackrel{(3.3)}{\sim} \frac{(1+\alpha)^2}{\sqrt{2\pi}\sigma_{\mathcal{S}}} \exp\left(-\frac{1}{2}\left((x \cdot (1 + \lambda_{\mathcal{S}})\frac{\sigma_1}{\sigma_{\mathcal{S}}} - z \cdot \frac{\sigma_3}{\sigma_{\mathcal{S}}})^2\right)\right) \end{aligned} \quad (3.10)$$

holds. Now, combining (3.9) and (3.10), we conclude that

$$\begin{aligned} \Psi_J &= \sum_{j=1}^K \sum_{n_1 \in J_j} \mathbb{P}[\mathcal{N}_1 = n_1] \mathbb{P}[\mathcal{S} = \nu - n_1 | \mathcal{N}_1 = n_1] \\ &\leq \frac{(1+\alpha)^3}{2\pi\sigma_{\mathcal{S}}} \sum_{j=1}^K \int_{I_j} \exp\left(-\frac{x^2}{2} - \frac{1}{2}\left((x \cdot (1 + \lambda_{\mathcal{S}})\frac{\sigma_1}{\sigma_{\mathcal{S}}} - z \cdot \frac{\sigma_3}{\sigma_{\mathcal{S}}})^2\right)\right) dx \\ &\leq \frac{1+4\alpha}{2\pi\sigma_{\mathcal{S}}} \int_{-\infty}^{\infty} \exp\left(-\frac{x^2}{2} - \frac{1}{2}\left((x \cdot (1 + \lambda_{\mathcal{S}})\frac{\sigma_1}{\sigma_{\mathcal{S}}} - z \cdot \frac{\sigma_3}{\sigma_{\mathcal{S}}})^2\right)\right) dx. \end{aligned} \quad (3.11)$$

Analogously, we derive the matching lower bound

$$\Psi_J \geq \frac{1-4\alpha}{2\pi\sigma_{\mathcal{S}}} \int_{-\infty}^{\infty} \exp\left(-\frac{x^2}{2} - \frac{1}{2}\left((x \cdot (1 + \lambda_{\mathcal{S}})\frac{\sigma_1}{\sigma_{\mathcal{S}}} - z \cdot \frac{\sigma_3}{\sigma_{\mathcal{S}}})^2\right)\right) dx. \quad (3.12)$$

Finally, combining (3.6), (3.8), (3.11), and (3.12), and remembering that $\mathbb{P}[\mathcal{N}_3 = \nu] = \Psi_J + \Psi_{J'} + \Psi_{J''}$, we obtain the assertion, because $\alpha > 0$ can be chosen arbitrarily small if n gets sufficiently large. \square

3.3 The Conditional Distribution of \mathcal{S}

Throughout this section, we keep the notation and the assumptions from Section 3.2. In addition, we let $G \subset V$ be a set of cardinality n_1 such that $|n_1 - \mu_1| \leq n^{0.6}$.

3.3.1 Outline

The goal of this section is to prove Proposition 3.3. Let us condition on the event that the largest component of H_1 is G . To analyse the conditional distribution of \mathcal{S} , we need to overcome the problem that in H_1 the edges in the set $V \setminus G$ do not occur independently anymore once we condition on G being the largest component of H_1 . However, we will see that this conditioning is “not very strong”. To this end, we shall compare \mathcal{S} with an

“artificial” random variable \mathcal{S}_G , which models the edges contained in $V \setminus G$ as mutually independent objects. To define \mathcal{S}_G , we set up random hypergraphs $H_{j,G}$, $j = 1, 2, 3$, in three “rounds” as follows.

- R1’.** The vertex set of $H_{1,G}$ is $V = \{1, \dots, n\}$, and each of the $\binom{n-n_1}{d}$ possible edges $e \subset V \setminus G$ is present in $H_{1,G}$ with probability p_1 independently.
- R2’.** Adding each possible edge $e \subset V \setminus G$ not present in $H_{1,G}$ with probability p_2 independently yields $H_{2,G}$.
- R3’.** Obtain $H_{3,G}$ from $H_{2,G}$ by including each possible edge e incident to both G and $V \setminus G$ with probability p_2 independently.

The process R1’–R3’ relates to the process R1–R4 from Section 3.2.1 as follows. While in H_1 the edges in $V \setminus G$ are mutually dependent, we have “artificially” constructed $H_{1,G}$ in such a way that the edges outside of G occur independently. Then, $H_{2,G}$ and $H_{3,G}$ are obtained similarly as H_2 and H_3 , namely by including further edges inside of $V \setminus G$ and crossing edges between G and $V \setminus G$ with probability p_2 . Letting \mathcal{S}_G denote the set of vertices in $V \setminus G$ that are reachable from G , the quantity $\mathcal{S}_G = |\mathcal{S}_G|$ now corresponds to \mathcal{S} . In contrast to R1–R4, the process R1’–R3’ completely disregards edges inside of G , because these do not affect \mathcal{S}_G . The following lemma, which we will prove in Section 3.3.3 shows that \mathcal{S}_G is indeed a very good approximation of \mathcal{S} , so that it suffices to study \mathcal{S}_G .

Lemma 3.5. *For any $\nu \in \mathbb{Z}$ we have $|\mathbb{P}[\mathcal{S} = \nu \mid \mathcal{N}_1 = n_1] - \mathbb{P}[\mathcal{S}_G = \nu]| \leq n^{-9}$.*

As a next step, we investigate the expectation of \mathcal{S}_G . While there is no need to compute $\mathbb{E}[\mathcal{S}_G]$ precisely, we do need that $\mathbb{E}[\mathcal{S}_G]$ depends on $n_1 - \mu_1$ linearly. The corresponding proof can be found in Section 3.3.4.

Lemma 3.6. *We have $\mathbb{E}[\mathcal{S}_G] = \mu_{\mathcal{S}} + \lambda_{\mathcal{S}}(n_1 - \mu_1) + o(\sqrt{n})$, where $\mu_{\mathcal{S}} = \Theta(n)$ and $\lambda_{\mathcal{S}} = \Theta(1)$ do not depend on n_1 .*

Furthermore, we need that the variance of \mathcal{S}_G is essentially independent of the precise value of n_1 . This will be proven in Section 3.3.5.

Lemma 3.7. *We have $\text{Var}[\mathcal{S}_G] = O(n)$. Moreover, if $G' \subset V$ is another set such that $|\mu_1 - |G'|| = o(n)$, then $|\text{Var}[\mathcal{S}_G] - \text{Var}[\mathcal{S}_{G'}]| = o(n)$.*

To show that \mathcal{S}_G satisfies a local limit theorem, the crucial step is to prove that for numbers s and t such that s is “close” to t the probabilities $\mathbb{P}[\mathcal{S}_G = s]$, $\mathbb{P}[\mathcal{S}_G = t]$ are “almost the same”. More precisely, the following lemma, proven in Section 3.3.2, holds.

Lemma 3.8. *For every $\alpha > 0$ there is $\beta > 0$ such that for all s, t satisfying $|s - \mathbb{E}[\mathcal{S}_G]|, |t - \mathbb{E}[\mathcal{S}_G]| \leq n^{0.6}$ and $|s - t| \leq \beta n^{1/2}$ we have*

$$(1 - \alpha)\mathbb{P}[\mathcal{S}_G = s] - n^{-10} \leq \mathbb{P}[\mathcal{S}_G = t] \leq (1 + \alpha)\mathbb{P}[\mathcal{S}_G = s] + n^{-10}.$$

Moreover, there is a constant $C > 0$ such that $\mathbb{P}[\mathcal{S}_G = s] \leq Cn^{-1/2}$ for all integers s .

Letting $G_0 = \{1, \dots, \lceil \mu_1 \rceil\}$, we define $\sigma_{\mathcal{S}}^2 = \text{Var}[\mathcal{S}_{G_0}]$ and obtain a lower bound on $\sigma_{\mathcal{S}}$ as an immediate consequence of Lemma 3.8.

Corollary 3.9. *We have $\sigma_{\mathcal{S}} = \Omega(\sqrt{n})$.*

Proof. By Lemma 3.8 there exists a number $0 < \beta < 0.01$ independent of n such that for all integers s, t satisfying $|s - \mathbb{E}[\mathcal{S}_G]|, |t - \mathbb{E}[\mathcal{S}_G]| \leq \sqrt{n}$ and $|s - t| \leq \beta\sqrt{n}$ we have

$$\mathbb{P}[\mathcal{S}_G = t] \geq \frac{2}{3} \mathbb{P}[\mathcal{S}_G = s] - n^{-10}. \quad (3.13)$$

Set $\gamma = \beta^2/64$ and assume for contradiction that $\sigma_{\mathcal{S}}^2 < \gamma n/2$. Moreover, suppose that $G = G_0 = \{1, \dots, \lceil \mu_1 \rceil\}$. Then Chebyshev's inequality entails that

$$\mathbb{P}[|\mathcal{S}_G - \mathbb{E}[\mathcal{S}_G]| \geq \sqrt{\gamma n}] \leq \frac{1}{2}.$$

Hence, there exists an integer s such that $|s - \mathbb{E}[\mathcal{S}_G]| \leq \sqrt{\gamma n}$ and $\mathbb{P}[\mathcal{S}_G = s] \geq \frac{1}{2}(\gamma n)^{-\frac{1}{2}}$. Therefore, due to (3.13) we have $\mathbb{P}[\mathcal{S}_G = t] \geq \frac{1}{4}(\gamma n)^{-\frac{1}{2}}$ for all integers t such that $|s - t| \leq \beta\sqrt{n}$. Thus, recalling that $\gamma = \beta^2/64$, we obtain $1 \geq \mathbb{P}[|\mathcal{S}_G - s| \leq \beta\sqrt{n}] = \sum_{t: |t-s| \leq \beta\sqrt{n}} \mathbb{P}[\mathcal{S}_G = t] \geq \frac{\beta\sqrt{n}}{4\sqrt{\gamma n}} > 1$. This contradiction shows that $\sigma_{\mathcal{S}}^2 \geq \gamma n/2$. \square

Using the above estimates of the expectation and the variance of \mathcal{S}_G and invoking Stein's method once more, in Section 3.4 we will show that \mathcal{S}_G is asymptotically normal.

Lemma 3.10. *If $|n_1 - \mu_1| \leq n^{0.66}$, then $(\mathcal{S}_G - \mathbb{E}[\mathcal{S}_G])/\sigma_{\mathcal{S}}$ is asymptotically normal.*

Proof of Proposition 3.3. The first part of the proposition follows readily from Lemmas 3.5 and 3.6. Moreover, the second assertion follows from Lemma 3.8. Furthermore, we shall establish below that

$$\mathbb{P}[\mathcal{S}_G = s] \sim \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{(s - \mathbb{E}[\mathcal{S}])^2}{2\sigma_{\mathcal{S}}^2}\right) \quad \text{for any } s \text{ such that } |s - \mathbb{E}[\mathcal{S}_G]| = O(\sqrt{n}). \quad (3.14)$$

This claim implies the third part of the proposition. For $(s - \mathbb{E}[\mathcal{S}])^2 \sigma_{\mathcal{S}}^{-2} \sim (\mu_{\mathcal{S}} + \lambda_{\mathcal{S}}(n_1 - \mu_1))^2 \sigma_{\mathcal{S}}^{-2}$ by Lemma 3.6 and Corollary 3.9, and $\mathbb{P}[\mathcal{S} = s | \mathcal{N}_1 = n_1] \sim \mathbb{P}[\mathcal{S}_G = s]$ by Lemma 3.5.

To prove (3.14) let $\alpha > 0$ be arbitrarily small but fixed. Since $\sigma_{\mathcal{S}}^2 = \Theta(n)$ by Lemma 3.7 and Corollary 3.9, Lemma 3.8 entails that for a sufficiently small $\beta > 0$ and all s, t satisfying $|s - \mathbb{E}[\mathcal{S}_G]|, |t - \mathbb{E}[\mathcal{S}_G]| \leq n^{0.6}$ and $|s - t| \leq \beta\sigma_{\mathcal{S}}$ we have

$$(1 - \alpha)\mathbb{P}[\mathcal{S}_G = s] - n^{-10} \leq \mathbb{P}[\mathcal{S}_G = t] \leq (1 + \alpha)\mathbb{P}[\mathcal{S}_G = s] + n^{-10}. \quad (3.15)$$

Now, suppose that s is an integer such that $|s - \mathbb{E}[\mathcal{S}_G]| \leq O(\sqrt{n})$, and set $z = (s - \mathbb{E}[\mathcal{S}_G])/\sigma_{\mathcal{S}}$. Then Lemma 3.10 implies that

$$\mathbb{P}[|\mathcal{S}_G - s| \leq \beta\sigma_{\mathcal{S}}] \geq \frac{1 - \alpha}{\sqrt{2\pi}} \int_{z-\beta}^{z+\beta} \exp(-x^2/2) dx \geq (1 - 2\alpha) \frac{\beta}{\sqrt{2\pi}} \exp(-z^2/2), \quad (3.16)$$

provided that β is small enough. Furthermore, (3.15) yields that

$$\begin{aligned}\mathbb{P}[|\mathcal{S}_G - s| \leq \beta\sigma_{\mathcal{S}}] &= \sum_{t:|t-s|\leq\beta\sigma_{\mathcal{S}}} \mathbb{P}[\mathcal{S}_G = t] \leq \beta\sigma_{\mathcal{S}}((1+\alpha)\mathbb{P}[\mathcal{S}_G = s] + n^{-10}) \\ &\leq (1+\alpha)\beta\sigma_{\mathcal{S}}\mathbb{P}[\mathcal{S}_G = s] + n^{-9},\end{aligned}\tag{3.17}$$

because $\sigma_{\mathcal{S}} = O(\sqrt{n})$ by Lemma 3.7. Combining (3.16) and (3.17), we conclude that

$$\mathbb{P}[\mathcal{S}_G = s] \geq \frac{1-2\alpha}{1+\alpha} \cdot \frac{1}{\sqrt{2\pi}\sigma_{\mathcal{S}}} \exp(-z^2/2) - n^{-9} \geq \frac{1-4\alpha}{\sqrt{2\pi}\sigma_{\mathcal{S}}} \exp\left(-\frac{(s - \mathbb{E}[\mathcal{S}_G])^2}{2\sigma_{\mathcal{S}}^2}\right).$$

Since analogous arguments yield the matching upper bound

$$\mathbb{P}[\mathcal{S}_G = s] \leq \frac{1+4\alpha}{\sqrt{2\pi}\sigma_{\mathcal{S}}} \exp\left(-\frac{(s - \mathbb{E}[\mathcal{S}_G])^2}{2\sigma_{\mathcal{S}}^2}\right),$$

and because $\alpha > 0$ may be chosen arbitrarily small, we obtain (3.14). \square

Next we will prove Lemma 3.8 which provides the central locality argument while the more technical proofs of Lemma 3.5, 3.6 and 3.7 are deferred to the end of this section.

3.3.2 Locality of \mathcal{S}_G

In this section we will prove Lemma 3.8. Since the assertion is symmetric in s and t , it suffices to prove that $\mathbb{P}[\mathcal{S}_G = s] \leq (1-\alpha)^{-1}\mathbb{P}[\mathcal{S}_G = s] + n^{-10}$. Let $\mathcal{F} = E(H_{3,G}) \setminus E(H_{2,G})$ be the (random) set of edges added during **R3'**. We split \mathcal{F} into three subsets: let \mathcal{F}_1 consist of all $e \in \mathcal{F}$ such that either $|e \setminus G| \geq 2$ or e contains a vertex that belongs to a component of $V \setminus G$ of order ≥ 2 . Moreover, \mathcal{F}_2 is the set of all edges $e \in \mathcal{F} \setminus \mathcal{F}_1$ that contain a vertex of $V \setminus G$ that is also contained in some other edge $e' \in \mathcal{F}_1$. Let $H'_{2,G}$ denote $H_{2,G}$ with the edges from \mathcal{F}_1 and \mathcal{F}_2 added. Finally, $\mathcal{F}_3 = \mathcal{F} \setminus (\mathcal{F}_1 \cup \mathcal{F}_2)$; thus, all edges $e \in \mathcal{F}_3$ connect $d-1$ vertices in G with a vertex $v \in V \setminus G$ that is isolated in $H'_{2,G}$, see Figure 3.1 for an example.

As a next step, we decompose \mathcal{S}_G into two contributions corresponding to $\mathcal{F}_1 \cup \mathcal{F}_2$ and \mathcal{F}_3 . More precisely, we let $\mathcal{S}_G^{\text{big}}$ be the number of vertices in $V \setminus G$ that are reachable from G in $H_{2,G} + \mathcal{F}_1 + \mathcal{F}_2$ and set $\mathcal{S}_G^{\text{iso}} = \mathcal{S}_G - \mathcal{S}_G^{\text{big}}$. Hence, if we let \mathcal{W} signify the set of all isolated vertices of $H_{2,G} + \mathcal{F}_1 + \mathcal{F}_2$ in the set $V \setminus G$, then $\mathcal{S}_G^{\text{iso}}$ equals the number of vertices in \mathcal{W} that get attached to G via the edges in \mathcal{F}_3 .

We can determine the distribution of $\mathcal{S}_G^{\text{iso}}$ precisely. For if $v \in \mathcal{W}$, then each edge e containing v and exactly $d-1$ vertices of G is present with probability p_2 independently. Therefore, the probability that v gets attached to G is $1 - (1-p_2)^{\binom{n_1}{d-1}}$. In fact, these events occur independently for all $v \in \mathcal{W}$. Consequently,

$$\mathcal{S}_G^{\text{iso}} = \text{Bi}\left(|\mathcal{W}|, 1 - (1-p_2)^{\binom{n_1}{d-1}}\right), \quad \mu_{\text{iso}} = \mathbb{E}[\mathcal{S}_G^{\text{iso}}] = |\mathcal{W}|(1 - (1-p_2)^{\binom{n_1}{d-1}}) = \Omega(|\mathcal{W}|),\tag{3.18}$$

where the last equality sign follows from the fact that $p_2 \sim \varepsilon p_1 = \Theta(n^{1-d})$.

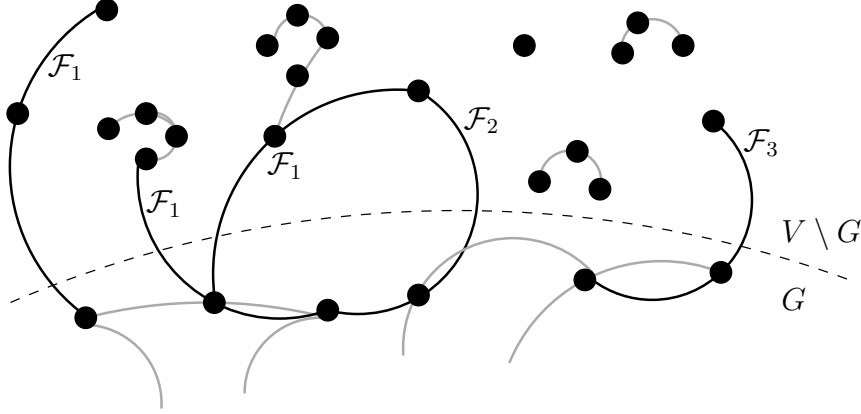


Figure 3.1: The three kinds of edges (black) which attach small components to G . The edges of $H_{2,G}$ are depicted in grey.

Hence, $\mathcal{S}_G = \mathcal{S}_G^{\text{big}} + \mathcal{S}_G^{\text{iso}}$ features a contribution that satisfies a local limit theorem, namely the binomially distributed $\mathcal{S}_G^{\text{iso}}$. Thus, to establish the locality of \mathcal{S}_G (i.e., Lemma 3.8), we are going to prove that \mathcal{S}_G “inherits” the locality of $\mathcal{S}_G^{\text{iso}}$. To this end, we need to bound $|\mathcal{W}|$, thereby estimating $\mu_{\text{iso}} = \mathbb{E}[\mathcal{S}_G^{\text{iso}}]$.

Lemma 3.11. *We have $\mathbb{P}\left[|\mathcal{W}| \geq \frac{1}{2}(n - n_1) \exp(-c)\right] \geq 1 - n^{-10}$.*

The proof of Lemma 3.11 is just a standard application of Azuma’s inequality, cf. Section 3.3.6.

Further, let M be the set of all triples (H, F_1, F_2) such that

M1. $\mathbb{P}[\mathcal{S}_G = s | H_{2,G} = H, \mathcal{F}_1 = F_1, \mathcal{F}_2 = F_2] \geq n^{-11}$, and

M2. given that $H_{2,G} = H$, $\mathcal{F}_1 = F_1$, and $\mathcal{F}_2 = F_2$, the set \mathcal{W} has size $\geq \frac{1}{2}(n - n_1) \exp(-c) = \Omega(n)$.

Lemma 3.12. *If $|s - t| \leq \beta\sqrt{n}$ for some small enough $\beta = \beta(\alpha) > 0$, then*

$$\mathbb{P}[\mathcal{S}_G = t | (H_{2,G}, \mathcal{F}_1, \mathcal{F}_2) \in M] \geq (1 - \alpha) \mathbb{P}[\mathcal{S}_G = s | (H_{2,G}, \mathcal{F}_1, \mathcal{F}_2) \in M].$$

Proof. Let $(H, F_1, F_2) \in M$, and let b be the value of $\mathcal{S}_G^{\text{big}}$ given that $H_{2,G} = H$, $\mathcal{F}_1 = F_1$ and $\mathcal{F}_2 = F_2$. Then given that this event occurs, we have $\mathcal{S}_G = s$ iff $\mathcal{S}_G^{\text{iso}} = s - b$. As $(H, F_1, F_2) \in M$, we conclude that

$$\begin{aligned} \mathbb{P}[\mathcal{S}_G = s | H_{2,G} = H, \mathcal{F}_1 = F_1, \mathcal{F}_2 = F_2] &= \mathbb{P}\left[\text{Bi}\left(|\mathcal{W}|, 1 - (1 - p_2)^{\binom{n_1}{d-1}}\right) = s - b\right] \\ &\stackrel{\text{M1}}{\geq} n^{-11}. \end{aligned}$$

Therefore, the Chernoff bound (1.3) implies that $|s - b - \mu_{\text{iso}}| \leq n^{0.6}$. Furthermore, since we assume that $|t - s| \leq \beta n^{1/2}$ for some small $\beta = \beta(\alpha) > 0$ and as $\mu_{\text{iso}} =$

$|\mathcal{W}|(1 - (1 - p_2)^{\binom{n_1}{d-1}}) \geq \Omega(n)$ due to **M2**, Proposition 1.1 entails that

$$\mathbb{P} \left[\text{Bi} \left(|\mathcal{W}|, 1 - (1 - p_2)^{\binom{n_1}{d-1}} \right) = t - b \right] \geq (1 - \alpha) \mathbb{P} \left[\text{Bi} \left(|\mathcal{W}|, 1 - (1 - p_2)^{\binom{n_1}{d-1}} \right) = s - b \right].$$

Thus, the assertion follows from (3.18). \square

Proof of Lemma 3.8. By Lemmas 3.11 and 3.12, we have

$$\begin{aligned} \mathbb{P}[\mathcal{S}_G = s] &\leq \mathbb{P}[\mathcal{S}_G = s | H'_{2,G} \notin M] \mathbb{P}[H'_{2,G} \notin M] + (1 - \alpha)^{-1} \mathbb{P}[\mathcal{S}_G = t] \\ &\stackrel{\mathbf{M1}, \mathbf{M2}}{\leq} n^{-11} + \mathbb{P}[|\mathcal{W}| = o(n)] + (1 - \alpha)^{-1} \mathbb{P}[\mathcal{S}_G = t] \\ &\leq (1 - \alpha)^{-1} \mathbb{P}[\mathcal{S}_G = t] + n^{-10}, \end{aligned}$$

as claimed. \square

3.3.3 Approximating \mathcal{S} via \mathcal{S}_G

This section contains the proof of Lemma 3.5. Let \mathcal{L}_G signify the event that G is the largest component of H_1 . Given that \mathcal{L}_G occurs, the edges in $H_3 - G$ do not occur independently anymore. For if \mathcal{L}_G occurs, then $H_1 - G$ does not contain a component on more than $|G|$ vertices. Nonetheless, the following lemma shows that if $E \subset \mathcal{E}(V) \setminus \mathcal{E}(G)$ is a set of edges such that the hypergraph $H(E) = (V, E \cap \mathcal{E}(V \setminus G))$ does not feature a “big” component, then the dependence of the edges is very small. In other words, the probability that the edges E are present in H_3 is very close to the probability that these edges are present in the “artificial” model $H_{3,G}$, in which edges occur independently.

Lemma 3.13. *For any set $E \subset \mathcal{E}(V) \setminus \mathcal{E}(G)$ such that $\mathcal{N}(H(E)) \leq \ln^2 n$ we have*

$$\mathbb{P}[E(H_3) \setminus \mathcal{E}(G) = E | \mathcal{L}_G] = (1 + O(n^{-10})) \mathbb{P}[E(H_{3,G}) = E].$$

Before getting down to the proof of Lemma 3.13, we first show how it implies Lemma 3.5. As a first step, we derive that it is actually quite unlikely that either $H_3 - G$ or $H_{3,G} - G$ features a component on $\geq \ln^2 n$ vertices.

Corollary 3.14. *We have*

$$\mathbb{P}[\mathcal{N}(H_3 - G) > \ln^2 n | \mathcal{L}_G], \mathbb{P}[\mathcal{N}(H_{3,G} - G) > \ln^2 n] = O(n^{-10}).$$

Proof. Theorem 1.2 implies that $\mathbb{P}[\mathcal{N}(H_{3,G} - G) > \ln^2 n] = O(n^{-10})$, because $H_{3,G}$ simply is a random hypergraph $H_d(n - n_1, p)$, and $\binom{n - n_1}{d-1} p \sim \binom{n - \mu_1}{d-1} p < (d - 1)^{-1}$ by (1.4). Hence, Lemma 3.13 yields that

$$\mathbb{P}[\mathcal{N}(H_3 - G) \leq \ln^2 n | \mathcal{L}_G] \geq (1 - O(n^{-10})) \mathbb{P}[\mathcal{N}(H_{3,G} - G) \leq \ln^2 n] \geq 1 - O(n^{-10}).$$

\square

Proof of Lemma 3.5. Let \mathcal{A}_s denote the set of all subsets $E \subset \mathcal{E}(V) \setminus \mathcal{E}(G)$ such that in the hypergraph (V, E) exactly s vertices in $V \setminus G$ are reachable from G . Moreover, let \mathcal{B}_s signify the set of all $E \in \mathcal{A}_s$ such that $\mathcal{N}(H(E)) \leq \ln^2 n$. Then

$$\mathbb{P}[\mathcal{S} = s | \mathcal{L}_G] = \mathbb{P}[E(H_3) \setminus \mathcal{E}(G) \in \mathcal{A}_s | \mathcal{L}_G], \text{ and } \mathbb{P}[\mathcal{S}_G = s] = \mathbb{P}[E(H_{3,G}) \in \mathcal{A}_s]. \quad (3.19)$$

Furthermore, by Corollary 3.14

$$\mathbb{P}[E(H_3) \setminus \mathcal{E}(G) \in \mathcal{A}_s \setminus \mathcal{B}_s | \mathcal{L}_G] \leq \mathbb{P}[\mathcal{N}(H_3 - G) > \ln^2 n | \mathcal{L}_G] = O(n^{-10}), \quad (3.20)$$

$$\mathbb{P}[E(H_{3,G}) \in \mathcal{A}_s \setminus \mathcal{B}_s] \leq \mathbb{P}[\mathcal{N}(H_{3,G} - G) > \ln^2 n] = O(n^{-10}). \quad (3.21)$$

Combining (3.19), (3.20), and (3.21), we conclude that

$$\begin{aligned} \mathbb{P}[\mathcal{S} = s | \mathcal{L}_G] &= \mathbb{P}[E(H_3) \setminus \mathcal{E}(G) \in \mathcal{B}_s | \mathcal{L}_G] + O(n^{-10}) \\ &\stackrel{\text{Lemma 3.13}}{=} \mathbb{P}[E(H_{3,G}) \in \mathcal{B}_s] + O(n^{-10}) = \mathbb{P}[\mathcal{S}_G = s] + O(n^{-10}), \end{aligned}$$

thereby completing the proof. \square

Thus, the remaining task is to prove Lemma 3.13. To this end, let $\mathcal{H}_1(E)$ denote the event that $\mathcal{E}(V \setminus G) \cap E(H_1) = E$. Moreover, let $\mathcal{H}_2(E)$ signify the event that $\mathcal{E}(V \setminus G) \cap E(H_2) \setminus E(H_1) = E$ (i.e., E is the set of edges added during **R2**). Further, let $\mathcal{H}_3(E)$ be the event that $\mathcal{E}(G, V \setminus G) \cap E(H_3) = E$ (i.e., E consists of all edges added by **R3**). In addition, define events $\mathcal{H}_{1,G}(E)$, $\mathcal{H}_{2,G}(E)$, $\mathcal{H}_{3,G}(E)$ analogously, with H_1 , H_2 , H_3 replaced by $H_{1,G}$, $H_{2,G}$, $H_{3,G}$. Finally, let \mathcal{C}_G denote the event that G is a component of H_1 . In order to prove Lemma 3.13, we establish the following.

Lemma 3.15. *Let $E_1 \subset \mathcal{E}(V \setminus G)$, $E_2 \subset \mathcal{E}(V \setminus G) \setminus E_1$, and $E_3 \subset \mathcal{E}(G, V \setminus G)$. Moreover, suppose that $\mathcal{N}(H(E_1)) \leq \ln^2 n$. Then*

$$\mathbb{P}\left[\bigwedge_{i=1}^3 \mathcal{H}_i(E_i) | \mathcal{L}_G\right] = (1 + O(n^{-10})) \mathbb{P}\left[\bigwedge_{i=1}^3 \mathcal{H}_{i,G}(E_i)\right].$$

Proof. Clearly,

$$\mathbb{P}\left[\bigwedge_{i=1}^3 \mathcal{H}_i(E_i) | \mathcal{L}_G\right] = \frac{\mathbb{P}[\mathcal{H}_2(E_2) \wedge \mathcal{H}_3(E_3) | \mathcal{L}_G \wedge \mathcal{H}_1(E_1)] \mathbb{P}[\mathcal{H}_1(E_1) \wedge \mathcal{L}_G]}{\mathbb{P}[\mathcal{L}_G]}. \quad (3.22)$$

Furthermore, since **R2** and **R3** add edges independently of the 1st round with probability p_2 , and because the same happens during **R2'** and **R3'**, we have

$$\mathbb{P}[\mathcal{H}_2(E_2) \wedge \mathcal{H}_3(E_3) | \mathcal{L}_G \wedge \mathcal{H}_1(E_1)] = \mathbb{P}[\mathcal{H}_{2,G}(E_2) \wedge \mathcal{H}_{3,G}(E_3) | \mathcal{H}_{1,G}(E_1)]. \quad (3.23)$$

Moreover, given that $\mathcal{H}_1(E_1)$ occurs, $H_1 - G$ has no component on more than $\ln^2 n$ vertices. Hence, G is the largest component of H_1 iff G is a component; that is, given that $\mathcal{H}_1(E_1)$ occurs, the events \mathcal{L}_G and \mathcal{C}_G are equivalent. Therefore, $\mathbb{P}[\mathcal{L}_G \wedge \mathcal{H}_1(E_1)] = \mathbb{P}[\mathcal{C}_G \wedge \mathcal{H}_1(E_1)]$. Further, whether or not G is a component of H_1 is independent of the

edges contained in $V \setminus G$, and thus $\mathbb{P}[\mathcal{C}_G \wedge \mathcal{H}_1(E_1)] = \mathbb{P}[\mathcal{C}_G] \mathbb{P}[\mathcal{H}_1(E_1)]$. Hence, as each edge in E_1 is present in H_1 as well as in $H_{1,G}$ with probability p_1 independently, we obtain

$$\mathbb{P}[\mathcal{L}_G \wedge \mathcal{H}_1(E_1)] = \mathbb{P}[\mathcal{C}_G] p_1^{|E_1|} (1 - p_1)^{\mathcal{E}(V \setminus G) - |E_1|} = \mathbb{P}[\mathcal{C}_G] \mathbb{P}[\mathcal{H}_{1,G}(E_1)]. \quad (3.24)$$

Combining (3.22), (3.23), and (3.24), we obtain

$$\mathbb{P} \left[\bigwedge_{i=1}^3 \mathcal{H}_i(E_i) | \mathcal{L}_G \right] = \frac{\mathbb{P}[\mathcal{C}_G]}{\mathbb{P}[\mathcal{L}_G]} \cdot \mathbb{P} \left[\bigwedge_{i=1}^3 \mathcal{H}_{i,G}(E_i) \right]. \quad (3.25)$$

Since by Theorem 1.2 with probability $\geq 1 - n^{-10}$ the random hypergraph $H_1 = H_d(n, p_1)$ has precisely one component of order $\Omega(n)$, we get $\frac{\mathbb{P}[\mathcal{C}_G]}{\mathbb{P}[\mathcal{L}_G]} = 1 + O(n^{-10})$. Hence, (3.25) implies the assertion. \square

Proof of Lemma 3.13. For any set $E \subset \mathcal{E}(V) \setminus \mathcal{E}(G)$ let $\mathcal{F}(E)$ denote the set of all decompositions (E_1, E_2, E_3) of E into three disjoint sets such that $E_1, E_2 \subset \mathcal{E}(V \setminus G)$ and $E_3 \subset \mathcal{E}(G, V \setminus G)$. If $\mathcal{N}(H(e)) \leq \ln^2 n$, then Lemma 3.15 implies that

$$\begin{aligned} \mathbb{P}[E(H_3) \setminus \mathcal{E}(G) = E | \mathcal{L}_G] &= \sum_{(E_1, E_2, E_3) \in \mathcal{F}(E)} \mathbb{P} \left[\bigwedge_{i=1}^3 \mathcal{H}_i(E_i) | \mathcal{L}_G \right] \\ &= (1 + O(n^{-10})) \sum_{(E_1, E_2, E_3) \in \mathcal{F}(E)} \mathbb{P} \left[\bigwedge_{i=1}^3 \mathcal{H}_{i,G}(E_i) \right] \\ &= (1 + O(n^{-10})) \mathbb{P}[E(H_{3,G}) = E], \end{aligned}$$

as claimed. \square

3.3.4 The Expectation of \mathcal{S}_G

The proof of Lemma 3.6 follows. Recall that \mathcal{S}_G signifies the set of all vertices $v \in V \setminus G$ that are reachable from G in $H_{3,G}$, so that $\mathcal{S}_G = |\mathcal{S}_G|$. Letting \mathcal{C}_v denote the component of $H_{2,G}$ that contains $v \in V$, we have

$$\mathbb{E}[\mathcal{S}_G] = \sum_{v \in V \setminus G} \mathbb{P}[v \in \mathcal{S}_G] = \sum_{v \in V \setminus G} \sum_{k=1}^{n-n_1} \mathbb{P}[v \in \mathcal{S}_G | |\mathcal{C}_v| = k] \mathbb{P}[|\mathcal{C}_v| = k] \quad (3.26)$$

Since $H_{2,G}$ is just a random hypergraph $H_d(n - n_1, p)$, and because $\binom{n-n_1}{d-1} p \sim \binom{n-\mu_1}{d-1} p < (d-1)^{-1}$ by (1.4), Theorem 1.2 entails that $\mathcal{N}(H_{2,G}) \leq \ln^2 n$ with probability $\geq 1 - n^{-10}$. Therefore, (3.26) yields

$$\mathbb{E}[\mathcal{S}_G] = o(1) + \sum_{v \in V \setminus G} \sum_{k=1}^{\ln^2 n} \mathbb{P}[v \in \mathcal{S}_G | |\mathcal{C}_v| = k] \mathbb{P}[|\mathcal{C}_v| = k]. \quad (3.27)$$

To estimate $\mathbb{P}[v \in S_G | |\mathcal{C}_v| = k]$, let $z = z(n_1) = (n_1 - \mu_1)/\sigma_1$,

$$\xi_0 = \exp\left(-p_2\left(\binom{n-1}{d-1} - \binom{n-\mu_1}{d-1}\right)\right), \text{ and } \xi(z) = \xi_0\left(1 + z\sigma_1 p_2 \binom{n-\mu_1}{d-2}\right).$$

Additionally, let $\zeta(z) = \binom{n-n_1}{d-1}p \sim \binom{n-\mu_1}{d-1}p - z\sigma_1 \binom{n-\mu_1}{d-2}p$.

Lemma 3.16. *For all $1 \leq k \leq \ln^2 n$ we have $\mathbb{P}[v \in S_G | |\mathcal{C}_v| = k] = 1 - \xi(z)^k + O(n^{-1} \cdot \text{polylog } n)$.*

Proof. Suppose that $|\mathcal{C}_v| = k$ but $v \notin S_G$. This is the case iff in $H_{3,G}$ there occurs no edge that is incident to both G and \mathcal{C}_v . Letting $\mathcal{E}(G, \mathcal{C}(v))$ denote the set of all possible edges connecting G and \mathcal{C}_v , we shall prove below that

$$\begin{aligned} |\mathcal{E}(G, \mathcal{C}_v)| &= k \left(\binom{n}{d-1} - \binom{n-\mu_1}{d-1} + \frac{z\sigma_1}{d-1} \binom{n-\mu_1}{d-2} \right) \\ &\quad + O(n^{d-2} \cdot \text{polylog } n) = O(n^{d-1} \cdot \text{polylog } n). \end{aligned} \quad (3.28)$$

By construction every edge in $\mathcal{E}(G, \mathcal{C}_v)$ occurs in $H_{3,G}$ with probability p_2 independently. Therefore,

$$\begin{aligned} \mathbb{P}[v \notin S_G | |\mathcal{C}_v| = k] &= (1 - p_2)^{|\mathcal{E}(G, \mathcal{C}_v)|} \\ &= (1 + O(n^{-1} \cdot \text{polylog } n)) \exp(-p_2 |\mathcal{E}(G, \mathcal{C}_v)|) \\ &\stackrel{(3.28)}{=} (1 + O(n^{-1} \cdot \text{polylog } n)) \xi(z)^k, \end{aligned}$$

hence the assertion follows.

Thus, the remaining task is to prove (3.28). As a first step, we show that

$$|\mathcal{E}(G, \mathcal{C}_v)| = \binom{n}{d} - \binom{n-k}{d} - \binom{n-n_1}{d} + \binom{n-n_1-k}{d}. \quad (3.29)$$

For there are $\binom{n}{d}$ possible edges in total, among which $\binom{n-k}{d}$ contain no vertex of \mathcal{C}_v , $\binom{n-n_1}{d}$ contain no vertex of G , and $\binom{n-n_1-k}{d}$ contain neither a vertex of \mathcal{C}_v nor of G ; thus, (3.29) follows from the inclusion/exclusion formula. Furthermore, as $k = O(\text{polylog } n)$, we have $\binom{n}{d} - \binom{n-k}{d} = (1 + O(n^{-1} \cdot \text{polylog } n))k \binom{n}{d-1}$ and $\binom{n-n_1}{d} - \binom{n-n_1-k}{d} = (1 + O(n^{-1} \cdot \text{polylog } n))k \binom{n-n_1}{d-1}$. Thus (3.29) yields

$$|\mathcal{E}(G, \mathcal{C}(v))| = (1 + O(n^{-1} \cdot \text{polylog } n))k \left(\binom{n}{d-1} - \binom{n-n_1}{d-1} \right). \quad (3.30)$$

As $n_1 = \mu_1 + z\sigma_1$, we have $\binom{n-n_1}{d-1} = \binom{n-\mu_1}{d-1} - z\sigma_1 \binom{n-\mu_1}{d-2} + O(n^{d-2} \cdot \text{polylog } n)$, so that (3.28) follows from (3.30). \square

Let $q(\zeta, \xi) = \sum_{k=1}^{\infty} q_k(\zeta) \xi^k$ be the function from Proposition 1.3. Combining (3.27) with Proposition 1.3 and Lemma 3.16, we conclude that

$$\mathbb{E}[\mathcal{S}_G] = o(n^{1/2}) + q((n - n_1)p, \xi(z))(n - n_1) = o(n^{1/2}) + q(\zeta(z), \xi(z))(n - n_1). \quad (3.31)$$

Since q is differentiable (cf. Proposition 1.3), we let $\Delta_\zeta = \frac{\partial q}{\partial \zeta}(\zeta(0), \xi(0))$ and $\Delta_\xi = \frac{\partial q}{\partial \xi}(\zeta(0), \xi(0))$. As $\zeta(z) - \zeta(0), \xi(z) - \xi(0) = O(n^{-1/2})$, we get

$$\begin{aligned} q(\zeta(z), \xi(z)) - q(\zeta(0), \xi(0)) &= (\zeta(z) - \zeta(0))\Delta_\zeta + (\xi(z) - \xi(0))\Delta_\xi + o(n^{-1/2}) \\ &= z\sigma_1 \binom{n - \mu_1}{d - 2} (\xi_0 \Delta_\xi p_2 - \Delta_\zeta p) + o(n^{-1/2}). \end{aligned} \quad (3.32)$$

Finally, let $\mu_S = (n - \mu_1)q(\zeta(0), \xi(0))$ and

$$\lambda_S = q(\zeta(0), \xi(0)) - (d - 1)(\varepsilon \xi_0 \Delta_\xi - \Delta_\zeta) \binom{n - \mu_1}{d - 1} p.$$

Then combining (3.31) and (3.32), we see that $\mathbb{E}[\mathcal{S}_G] = \mu_S + z\sigma_1 \lambda_S + o(\sqrt{n})$, as desired.

3.3.5 The Variance of \mathcal{S}_G

Remember that S_G denotes the set of all “attached” vertices, and $N_{v,G}$ the order of the component of $v \in V \setminus G$ in the graph $H_{2,G}$.

The following lemma provides an asymptotic formula for $\text{Var}[\mathcal{S}_G]$.

Lemma 3.17. *Let $r_{G,i} = \mathbb{P}[N_{v,G} = i \wedge v \in S_G]$ and $\bar{r}_{G,i} = \mathbb{P}[N_{v,G} = i \wedge v \notin S_G]$ for any vertex $v \in V \setminus G$. Moreover, set $r_G = \sum_{i=1}^L r_{G,i}$, $R_G = \sum_{i=1}^L i r_{G,i}$, $\bar{R}_G = \sum_{i=1}^L i \bar{r}_{G,i}$ for $L = \lceil \ln^2 n \rceil$. In addition, let $\alpha_G = 1 - |G|/n$ and*

$$\Gamma_G = (1 - R_G)(R_G - r_G) + ((d - 1)c - 1)R_G^2 + r_G R_G + (d - 1)(1 - \alpha_G^{d-2})\varepsilon c \bar{R}_G^2 + \frac{1 - \alpha_G^{d-2}}{1 - \alpha_G^{d-1}} \bar{R}_G. \quad (3.33)$$

Then $\text{Var}[\mathcal{S}_G] \sim \alpha_G^2 \Gamma_G n + \alpha_G r_G (1 - r_G) n$.

Before we get down to the proof of Lemma 3.17, we observe that it implies Lemma 3.7.

Proof of Lemma 3.7. By Theorem 1.2 part 2 together with Lemma 3.5 we know that with probability at least $1 - n^{-8}$ there are no components of order $> \ln^2 n$ inside of $V \setminus G$. Let $q(\zeta, \xi) = \sum_{k=1}^{\infty} q_k(\zeta) \xi^k$ be the function from Proposition 1.3, and let $\xi(z)$ be as in Lemma 3.16. Then Proposition 1.3 and Lemma 3.16 entail that for all $v \in V \setminus G$

$$r_{G,i} = q_i \left(\binom{n - |G|}{d - 1} p \right) \xi((|G| - \mu_1)/\sigma_1), \quad \bar{r}_{G,i} \sim q_i \left(\binom{n - |G|}{d - 1} p \right) (1 - \xi((|G| - \mu_1)/\sigma_1)).$$

By (1.6) there exists a number $0 < \gamma < 1$ such that $q_i \left(\binom{n - |G|}{d - 1} p \right) \leq \gamma^i$. Since $0 \leq \xi((|G| - \mu_1)/\sigma_1) \leq 1$, this yields $r_{G,i}, \bar{r}_{G,i} \leq \gamma^i$. Hence, $R_G, \bar{R}_G = O(1)$, so that Lemma 3.17 implies $\text{Var}[\mathcal{S}_G] = O(n)$.

Finally, if $G' \subset V$ satisfies $||G'| - |G|| \leq n^{0.9}$, then $|(n-|G|)p - (n-|G'|)p| = O(|G| - |G'|)/n$, because $p = O(n^{1-d})$. Therefore, $|q_i((n-|G|)p) - q_i((n-|G'|)p)| = O(|G| - |G'|)/n$, because the function $\zeta \mapsto q_i(\zeta)$ is differentiable. Similarly, as $\xi(z) = \xi_0(1 + z\sigma_1 p_2(\frac{n-\mu_1}{d-2}))$ for some fixed $\xi_0 = \Theta(1)$, we have $|\xi((|G| - \mu_1)/\sigma_1) - \xi((|G'| - \mu_1)/\sigma_1)| = O(|G| - |G'|)/n$. Consequently, $|r_{G,i} - r_{G',i}| = O(|G| - |G'|)/n$ and $|\bar{r}_{G,i} - \bar{r}_{G',i}| = O(|G| - |G'|)/n$, and thus

$$|r_G - r_{G'}|, |R_G - R_{G'}|, |\bar{R}_G - \bar{R}_{G'}| = O(|G| - |G'|)/n = O(n^{-0.1}).$$

Hence, Lemma 3.17 implies that $|\text{Var}[\mathcal{S}_G] - \text{Var}[\mathcal{S}_{G'}]| = o(n)$. \square

The remaining task is to establish Lemma 3.17. Keeping G fixed, we constantly omit the subscript G up to the end of this section (except when referring to $H_{3,G}$) in order to ease up the notation; thus, we write α instead of α_G etc. As a first step, we compute $\mathbb{P}[(v, w \in S) - r^2]$. Setting

$$\begin{aligned} S_1 &= \sum_{i,j=1}^L (\mathbb{P}[N_w = j \wedge y \in S | w \notin C_v, N_v = i, v \in S] - \mathbb{P}[N_w = j \wedge w \in S]) \\ &\quad \times \mathbb{P}[w \notin C_v | N_v = i, v \in S] \mathbb{P}[N_v = i \wedge v \in S], \\ S_2 &= (1 - r) \sum_{i=1}^L \mathbb{P}[w \in C_v | N_v = i, v \in S] \mathbb{P}[N_v = i \wedge v \in S], \end{aligned}$$

we have $\mathbb{P}[(v, w \in S) - r^2] = S_1 + S_2$.

To compute S_2 , observe that whether $w \in C_v$ depends only on N_v , but not on the event $v \in S$. Therefore, $\mathbb{P}[w \in C_v | N_v = i, v \in S] = \mathbb{P}[w \in C_v | N_v = i] = \frac{\binom{n-2}{x-2} \binom{n-1}{i-1}^{-1}}{\frac{i-1}{n-1}}$, because given that $N_v = i$, there are $\binom{n_0-1}{i-1}$ ways to choose the set $C_v \subset V \setminus G$, while there are $\binom{n_0-2}{i-2}$ ways to choose C_v in such a way that $w \in C_v$. As a consequence,

$$S_2 \sim \frac{1-r}{n-1} \sum_{i=1}^L (i-1) \mathbb{P}[N_v = i \wedge v \in S] = \frac{1-r}{n-1} (R-r).$$

With respect to S_1 , we let

$$\begin{aligned} P_1(i, j) &= \mathbb{P}[N_w = j | w \notin C_v, N_v = i], \\ P_2(i, j) &= \mathbb{P}[w \in S | N_w = j, w \notin C_v, N_v = i, v \in S], \end{aligned}$$

so that

$$\begin{aligned} S_1 &= \sum_{i,j} (P_1(i, j) P_2(i, j) - \mathbb{P}[N_w = j \wedge w \in S]) \mathbb{P}[w \notin C_v | N_v = i, v \in S] \mathbb{P}[N_v = i \wedge v \in S] \\ &\sim \sum_{i,j} (P_1(i, j) P_2(i, j) - \mathbb{P}[N_w = j] \mathbb{P}[w \in S | N_w = j]) \mathbb{P}[N_v = i \wedge v \in S]. \end{aligned}$$

Lemma 3.18. *We have $P_1(i, j) \mathbb{P}[N_w = j]^{-1} = 1 + \frac{((d-1)c-1)ij+i}{n} + O(n^{-2} \cdot \text{polylog } n)$.*

Proof. This argument is similar to the one used in the proof of Lemma 41 in Coja-Oghlan et al. [2006]. Remember that if we restrict our view on $H_{3,G}$ to the set $V \setminus G$ the hypergraph is similar to a $H_d(n - n_1, p)$. In order to estimate S_1 , we observe that

$$\mathbb{P}[N_w = j \text{ in } H_d(n - n_1, p) \mid N_v = i, w \notin C_v] = \mathbb{P}[N_w = j \text{ in } H_d(n - n_1, p) \setminus C_v]. \quad (3.34)$$

Given that $N_v = i$, $H_d(n, p) \setminus C_v$ is distributed as a random hypergraph $H_d(n - n_1 - i, p)$. Hence, the probability that $N_w = j$ in $H_d(n, p) \setminus C_v$ equals the probability that a given vertex of $H_d(n - n_1 - i, p)$ belongs to a component of order j . Therefore, we can compare $\mathbb{P}[N_w = j \text{ in } H_d(n - n_1, p) \setminus C_v]$ and $\mathbb{P}[N_w = j \text{ in } H_d(n - n_1, p)]$ as follows: in $H_d(n - n_1 - i, p)$ there are $\binom{n - n_1 - i - 1}{j - 1}$ ways to choose the set $C_w \setminus \{j\}$. Moreover, there are $\binom{n - n_1 - i}{d} - \binom{n - n_1 - i - j}{d} - \binom{j}{d}$ possible edges connecting the chosen set C_w with $V \setminus C_w$, and as C_w is a component, none of these edges is present. Since each such edge is present with probability p independently, the probability that there is no edge connecting C_w and $V \setminus C_w$ equals

$$(1 - p)^{\binom{n - n_1 - i}{d} - \binom{n - n_1 - i - j}{d} - \binom{j}{d}}.$$

By comparison, in $H_d(n - n_1, p)$ there are $\binom{n - n_1 - 1}{j - 1}$ ways to choose the vertex set of C_w . Further, there are $\binom{n - n_1}{d} - \binom{n - n_1 - j}{d} - \binom{j}{d}$ possible edges connecting C_w and $V \setminus C_w$, each of which is present with probability p independently. Thus, letting $\gamma = \binom{n - n_1 - i}{d} - \binom{n - n_1 - i - j}{d} - \left(\binom{n - n_1}{d} - \binom{n - n_1 - j}{d} \right)$, we obtain

$$\frac{\mathbb{P}[N_w = j \text{ in } H_d(n - n_1, p) \setminus C_v]}{\mathbb{P}[N_w = j \text{ in } H_d(n - n_1, p)]} = \binom{n - n_1 - i - 1}{j - 1} \binom{n - n_1 - 1}{j - 1}^{-1} (1 - p)^\gamma. \quad (3.35)$$

Concerning the quotient of the binomial coefficients, we have

$$\binom{n - n_1 - i - 1}{j - 1} \binom{n - n_1 - 1}{j - 1}^{-1} = \exp \left(-\frac{x(y - 1)}{n - n_1} + O(n^{-2} \cdot \text{polylog } n) \right). \quad (3.36)$$

Moreover, $\gamma = \binom{n - n_1}{d} \left(\frac{(n - n_1 - i)_d + (n - n_1 - j)_d - (n - n_1 - i - j)_d}{(n - n_1)_d} - 1 \right)$. Expanding the falling factorials, we get

$$\begin{aligned} \gamma &= \binom{n - n_1}{d} \left(\frac{\binom{d}{2}(i^2 + j^2 - (i + j)^2)}{(n - n_1)^2} + O(n^{-3} \cdot \text{polylog } n) \right) \\ &= -\binom{n - n_1}{d - 2} ij + O(n^{d-3} \cdot \text{polylog } n). \end{aligned} \quad (3.37)$$

Plugging (3.36) and (3.37) into (3.35), we obtain

$$\begin{aligned}
 & \frac{\mathbb{P}[N_w = j \text{ in } H_d(n - n_1, p) \setminus C_v]}{\mathbb{P}[N_w = j \text{ in } H_d(n - n_1, p)]} \\
 &= \exp\left(-\frac{i(j-1)}{n-n_1} + O\left(n^{-2} \cdot \text{polylog } n\right)\right) (1-p)^{-\binom{n-n_1}{d-2}ij + O(n^{d-3} \cdot \text{polylog } n)} \\
 &= \exp\left(-\frac{i(j-1)}{n-n_1} + \left(\frac{n-n_1}{d-2}\right)ijp + O\left(n^{-2} \cdot \text{polylog } n\right)\right) \\
 &= 1 + (n-n_1)^{-1}((d-1)c-1)ij + i + O\left(n^{-2} \cdot \text{polylog } n\right).
 \end{aligned}$$

Therefore, by (3.34)

$$\begin{aligned}
 & \mathbb{P}[N_w = j | N_v = i, w \notin C_v] - \mathbb{P}[N_w = j \text{ in } H_d(n - n_1, p)] \\
 &= \mathbb{P}[N_w = j \text{ in } H_d(n - n_1, p)] \left(n^{-1}(((d-1)c-1)ij + i) + O\left(n^{-2} \cdot \text{polylog } n\right) \right).
 \end{aligned} \tag{3.38}$$

□

Lemma 3.19. *Setting $\gamma_1 = \frac{1-\alpha^{d-2}}{\mathbb{P}[v \in S | N_v = i](1-\alpha^{d-1})}$ and $\gamma_2 = (d-1)(1-\alpha^{d-2})\varepsilon c$, we have $P_2(i, j) - \mathbb{P}[w \in S | N_w = j] = n^{-1}\mathbb{P}[w \notin S | N_w = j](j\gamma_1 - ij\gamma_2) + O(n^{-2} \cdot \text{polylog } n)$.*

Proof. Let \mathcal{F} be the event that $N_w = j$, $w \notin C_v$, $N_v = i$, and $v \in S$. Moreover, let \mathcal{Q} be the event that in H_3 there exists an edge incident to the three sets C_v , C_w , and G simultaneously, so that $P_2(i, j) = \mathbb{P}[\mathcal{Q} | \mathcal{F}] + \mathbb{P}[w \in S | \neg \mathcal{Q}, \mathcal{F}] \mathbb{P}[\neg \mathcal{Q} | \mathcal{F}]$.

To bound $\mathbb{P}[w \in S | \neg \mathcal{Q}, \mathcal{F}] - \mathbb{P}[w \in S | N_w = j]$, we condition on the event that C_v and C_w are fixed disjoint sets of sizes i and j . Let Q' signify the probability that C_w is reachable from G in $H_{3,G}$, and let Q denote the probability that C_w is reachable from G in $H_{3,G}$, and that the event $\neg \mathcal{Q}$ occurs. Then Q' corresponds to $\mathbb{P}[w \in S | N_w = j]$ and Q to $\mathbb{P}[w \in S | \neg \mathcal{Q}, \mathcal{F}]$, so that our aim is to estimate $Q - Q'$. As there are $|\mathcal{E}(G, C_v)| - |\mathcal{E}(G, C_v, C_w)|$ possible edges that join C_v and G but avoid C_w , each of which is present in $H_{3,G}$ with probability p_2 independently, we have

$$Q = 1 - (1 - p_2)^{|\mathcal{E}(G, C_v)| - |\mathcal{E}(G, C_v, C_w)|}, \text{ while } Q' = 1 - (1 - p_2)^{|\mathcal{E}(G, C_w)|}.$$

Therefore,

$$\begin{aligned}
 Q - Q' &= (1 - p_2)^{|\mathcal{E}(G, C_w)|} \left(1 - (1 - p_2)^{-|\mathcal{E}(G, C_v, C_w)|} \right) \\
 &\sim (1 - Q')(1 - \exp(p_2 |\mathcal{E}(G, C_v, C_w)|)) \sim ij(Q' - 1) \left(\binom{n}{d-2} - \binom{n_0}{d-2} \right) p_2.
 \end{aligned}$$

As $\binom{n-1}{d-1}p_2 \sim \varepsilon c$, we thus get

$$\mathbb{P}[w \in S | \neg \mathcal{Q}, \mathcal{F}] - \mathbb{P}[w \in S | N_w = j] \sim ij(\mathbb{P}[w \in S | N_w = j] - 1)(d-1)(1-\alpha^{d-2})\varepsilon cn^{-1}. \tag{3.39}$$

With respect to $\mathbb{P}[\mathcal{Q}|\mathcal{F}]$, we let \mathcal{K} signify the number of edges joining C_v and G . Given that \mathcal{F} occurs, \mathcal{K} is asymptotically Poisson with mean $\lambda_i = i \left(\binom{n}{d-1} - \binom{n_0}{d-1} \right) p_2 \sim i(1 - \alpha^{d-1})\varepsilon c$. Moreover, given that $\mathcal{K} = k$, the probability that one of these k edges hits C_w is $p^*(k) \sim \frac{k\mathcal{E}(G, C_v, C_w)}{\mathcal{E}(C_v, G)}$, and thus

$$p^*(k) \sim jk \left(\binom{n}{d-2} - \binom{n_0}{d-2} \right) \left(\binom{n}{d-1} - \binom{n_0}{d-1} \right)^{-1} \sim jk(d-1) \frac{1 - \alpha^{d-2}}{1 - \alpha^{d-1}}.$$

Consequently,

$$\mathbb{P}[\mathcal{Q}|\mathcal{F}] \sim \frac{\exp(-\lambda_i)}{1 - \exp(-\lambda_i)} \sum_{k \geq 1} \frac{jk\lambda_i^k}{k!} p^*(k) \sim \frac{j(1 - \alpha^{d-2})}{n(1 - \exp(-\lambda_i))(1 - \alpha^{d-1})}. \quad (3.40)$$

Combining (3.39) and (3.40), we obtain the assertion. \square

Thus,

$$\begin{aligned} nS_1 &\sim \sum_{i=1}^L \mathbb{P}[v \in S \wedge N_v = i] \\ &\cdot \sum_{j=1}^L (((d-1)c-1)ij + i) \mathbb{P}[w \in S \wedge N_w = j] + \mathbb{P}[w \notin S \wedge N_w = j](\gamma_1 j + \gamma_2 ij) \\ &= ((d-1)c-1) + R^2 + rR + \gamma_2 \bar{R}^2 + \sum_{i=1}^N \frac{1 - \alpha^{d-2}}{1 - \alpha^{d-1}} \mathbb{P}[N_v = i] \bar{R} \\ &= ((d-1)c-1)R^2 + rR + (d-1)(1 - \alpha^{d-2})\varepsilon c \bar{R}^2 + \frac{1 - \alpha^{d-2}}{1 - \alpha^{d-1}} \bar{R}. \end{aligned}$$

Hence, letting Γ be as defined by (3.33) we have $\mathbb{P}[v, w \in S] - \mathbb{P}[v \in S]\mathbb{P}[w \in S] \sim \Gamma/n$. Consequently, $\text{Var}[S] \sim \alpha\Gamma n + \alpha^2 r(1-r)n$.

3.3.6 The Number of Attached Isolated Vertices

Now we prove Lemma 3.11. The probability that a vertex $v \in V \setminus G$ is isolated in $H_{3,G}$ is at least $(1-p)^{\binom{n_1-1}{d-1}}(1-p_2)^{\binom{n}{d-1}} \sim \exp(-p\binom{n_1-1}{d-1} - \varepsilon p\binom{n}{d-1}) \geq \exp(-c)$. Therefore,

$$\mathbb{E}[|\mathcal{W}|] \geq (1 - o(1)) \exp(-c)(n - n_1). \quad (3.41)$$

To show that $|\mathcal{W}|$ is concentrated about its mean, we employ the following version of Azuma's inequality (cf. [Janson et al., 2000, p. 38]).

Lemma 3.20. *Let $\Omega = \prod_{i=1}^K \Omega_i$ be a product of probability spaces. Moreover, let $X : \Omega \rightarrow \mathbb{R}$ be a random variable that satisfies the following Lipschitz condition.*

$$\text{If two } k\text{-tuples } \omega = (\omega_i)_{1 \leq i \leq K}, \omega' = (\omega'_i)_{1 \leq i \leq K} \in \Omega \text{ differ only in their } j\text{'th components for some } 1 \leq j \leq K, \text{ then } |X(\omega) - X(\omega')| \leq 1. \quad (3.42)$$

Then $\mathbb{P}[|X - \mathbb{E}[X]| \geq t] \leq 2 \exp(-\frac{t^2}{2K})$, provided that $\mathbb{E}[X]$ exists.

Using Lemma 3.20, we shall establish the following.

Corollary 3.21. *Let Y be a random variable that maps the set of all d -uniform hypergraphs with vertex set V to $(0, n)$. Assume that Y satisfies the following condition.*

$$\text{Let } H \text{ be a hypergraph, and let } e \in \mathcal{E}(V). \text{ Then } |Y(H) - Y(H + e)|, |Y(H) - Y(H - e)| \leq 1. \quad (3.43)$$

Then $\mathbb{P}[|Y(H_{3,G}) - \mathbb{E}[Y(H_{3,G})]| \geq n^{0.66}] \leq \exp(-n^{0.01})$.

Proof. In order to apply Lemma 3.20, we need to decompose $H_{3,G}$ into a product $\prod_{i=1}^K \Omega_i$ of probability spaces. To this end, consider an arbitrary decomposition of the set $\mathcal{E}(V)$ of all possible edges into sets $\mathcal{E}_1 \cup \dots \cup \mathcal{E}_K$ so that $K \leq n$ and $\mathbb{E}[|E(H_{3,G}) \cap \mathcal{E}_j|] \leq n^{0.1}$ for all $1 \leq j \leq K$; such a decomposition exists, because the expected number of edges of $H_{3,G}$ is $\leq \binom{n}{d}p = O(n)$. Now, let Ω_e be a Bernoulli experiment with success probability p for each $e \in \mathcal{E}(V \setminus G)$, resp. with success probability p_2 for $e \in \mathcal{E}(G, V \setminus G)$. Then setting $\Omega_i = \prod_{e \in \mathcal{E}_i} \Omega_e$, we obtain a product decomposition $H_{3,G} = \prod_{i=1}^K \Omega_i$.

In addition, construct for each hypergraph H with vertex set V another hypergraph H^* by removing from H all edges $e \in \mathcal{E}_i$ such that $|E(H) \cap \mathcal{E}_i| \geq 4n^{0.1}$ ($1 \leq i \leq K$). Since $|E(H_{3,G}) \cap \mathcal{E}_i|$ is the sum of two binomially distributed variables, the Chernoff bound (1.3) implies that $\mathbb{P}[|E(H_{3,G}) \cap \mathcal{E}_i| \geq 4n^{0.1}] \leq \exp(-n^{0.05})$. As $K \leq n$, this entails

$$\mathbb{P}[H_{3,G} \neq H_{3,G}^*] \leq K \exp(-n^{0.05}) \leq \exp(-n^{0.04}), \text{ so that} \quad (3.44)$$

$$|\mathbb{E}[Y(H_{3,G})] - \mathbb{E}[Y(H_{3,G}^*)]| \leq 1 \quad [\text{because } 0 \leq Y \leq n]. \quad (3.45)$$

As a next step, we claim that $Y^*(H) = \frac{1}{4}n^{-0.1}Y(H^*)$ satisfies the Lipschitz condition (3.42). For by construction modifying (i.e., adding or removing) an arbitrary number of edges belonging to a single factor \mathcal{E}_i can affect at most $4n^{0.1}$ edges of H^* . Hence, (3.43) implies that $Y^*(H)$ satisfies (3.42). Therefore, Lemma 3.20 entails that

$$\begin{aligned} \mathbb{P}[|Y(H_{3,G}^*) - \mathbb{E}[Y(H_{3,G}^*)]| \geq n^{0.63}] &\leq \mathbb{P}[|Y^*(H_{3,G}) - \mathbb{E}[Y^*(H_{3,G})]| \geq n^{0.52}] \\ &\leq \exp(-n^{0.02}). \end{aligned} \quad (3.46)$$

Finally, combining (3.44), (3.45), and (3.46), we conclude that

$$\begin{aligned} \mathbb{P}[|Y(H_{3,G}) - \mathbb{E}[Y(H_{3,G})]| \geq n^{0.64}] &\leq \mathbb{P}[|Y^*(H) - \mathbb{E}[Y^*(H)]| \geq n^{0.63}] \\ &\quad + \mathbb{P}[H_{3,G} \neq H_{3,G}^*] \\ &\leq \exp(-n^{0.01}), \end{aligned}$$

thereby completing the proof. \square

Finally, since $|\mathcal{W}|/d$ satisfies (3.43), Lemma 3.11 follows from Corollary 3.21 and (3.41).

3.4 Central Limit Theorem for \mathcal{S}

In this section we will prove the properties **Y1**–**Y6** defined in Lemma 2.4 for the case of the normality of $|\mathcal{S}|$ where \mathcal{S} is the set of vertices which get attached to the giant in the second round of our two round exposure.

Consider a set $G \subset V$ of size n_1 . Let \mathcal{A} be the set of all subsets $\alpha \subset V \setminus G$ of size $|\alpha| \leq k$. Moreover, let $p_e = p$ for $e \subset V \setminus G$, $p_e = p_2$ for $e \in \mathcal{E}(G, V \setminus G)$, and $p_e = 0$ if $e \subset G$.

For $A \subseteq V$ and $A \cap \alpha = \emptyset$ set $I_\alpha^A = 1$ if α is a component of $H \setminus \mathcal{E}(A \cup G)$. Moreover, let $J_\alpha^A = 1$ if $(H \setminus \mathcal{E}(A)) \cap \mathcal{E}(G, \alpha) \neq \emptyset$. Further, let $K_\alpha^A = I_\alpha^A J_\alpha^A$ and $Y_\alpha^A = |\alpha| K_\alpha^A$. Then

$$\mathbb{P}[K_\alpha = 1] = \Omega(\mathbb{P}[I_\alpha = 1]). \quad (3.47)$$

Proof of Y1: The order of magnitude of $\mathbb{E}[Y]$ and $\text{Var}[Y]$ was already shown in Section 3.3. The proof of the rest of Y1 is the same as in the last section.

Proof of Y2: (2.12): Suppose that $K_\alpha = 1$. Then $I_\alpha = 1$, so that $H \setminus \mathcal{E}(G)$ has no α - β -edges. Hence, if also $K_\beta^\alpha = 1$, then β is a component of $H \setminus \mathcal{E}(G)$ as well. Thus, $K_\beta = 1$, so that $Y_\beta = Y_\beta^\alpha$.

(2.13): If $K_\alpha = 1$, then α is a component of $H \setminus \mathcal{E}(G)$. Since any two components of $H \setminus \mathcal{E}(G)$ are either disjoint or equal, we obtain $I_\beta = 0$, so that $Y_\beta = 0$ as well.

(2.14): To show that $Y_\gamma(Y_\beta - Y_\beta^\alpha) = 0$, assume that $K_\gamma = 1$. Then $I_\gamma = 1$, i.e., γ is a component of $H \setminus \mathcal{E}(G)$. Since $\beta \neq \gamma$ but $\beta \cap \gamma \neq \emptyset$, we conclude that $I_\beta = 0$. Furthermore, if γ is a component of $H \setminus \mathcal{E}(G)$, then γ is also a component of $H \setminus \mathcal{E}(G \cup \alpha)$, whence $I_\beta^\alpha = 0$. Consequently, $Y_\beta = Y_\beta^\alpha = 0$.

In order to prove that $Y_\gamma^\alpha(Y_\beta - Y_\beta^\alpha) = 0$, suppose that $K_\gamma^\alpha = 1$. Then $K_\gamma = 1$. Therefore, $I_\beta^\alpha = 0$, because the intersecting sets β, γ cannot both be components of $H \setminus \mathcal{E}(\alpha)$. Thus, we also have $I_\beta = 0$; for if β were a component of H , then β would also be a component of $H \setminus \mathcal{E}(\alpha)$. Hence, also in this case we obtain $Y_\beta = Y_\beta^\alpha = 0$.

Proof of Y3: Suppose that $K_\beta = 1$. Then $I_\beta = 1$, i.e., β is a component of $H \setminus \mathcal{E}(G)$. Then removing the edges \mathcal{E}_α from $H \setminus \mathcal{E}(G)$ may cause β to split into several components B_1, \dots, B_l . Thus, if $Y_\gamma^\beta > 0$ for some $\gamma \in \mathcal{A}$ such that $\gamma \cap \beta \neq \emptyset$, then γ is one of the components B_1, \dots, B_l . Since $l \leq |\beta| \leq k$, this implies that given $I_\beta = 1$ we have the bound

$$\sum_{\gamma: \gamma \cap \beta \neq \emptyset, \gamma \cap \alpha = \emptyset} Y_\gamma^\alpha \leq k^2.$$

Hence, we obtain **Y3**.

Lemma 3.22. *Let $0 \leq l, r \leq 2$, and let $\alpha_1, \dots, \alpha_l, \beta_1, \dots, \beta_r \in \mathcal{A}$ be pairwise disjoint. Moreover, let $A_1, \dots, A_r, B_1, \dots, B_r \subset V$ be sets such that $A_i \subset B_i \subset V \setminus \beta_i$ and $|B_i| \leq O(1)$ for all $1 \leq i \leq r$, and assume that $\bigcap_{i=1}^r B_i \setminus A_i = \emptyset$. Then*

$$\mathbb{P} \left[\bigwedge_{i=1}^l \bigwedge_{j=1}^r K_{\alpha_i} = 1 \wedge K_{\beta_j}^{A_j} \neq K_{\beta_j}^{B_j} \right] \leq O(n^{-r} \cdot \text{polylog } n) \prod_{j=1}^l \mathbb{P}[K_{\alpha_i} = 1] \prod_{j=1}^r \mathbb{P}[K_{\beta_j} = 1].$$

Similarly to Lemma 2.6 this lemma easily implies **Y4–Y6**.

Proof. Let $\tilde{p} = \mathbb{P} \left[\forall i, j : K_{\alpha_i} = 1 \wedge K_{\beta_j}^{A_j} \neq K_{\beta_j}^{B_j} \right]$. If $K_{\beta_j}^{A_j} \neq K_{\beta_j}^{B_j}$, then either $I_{\beta_j}^{A_j} \neq I_{\beta_j}^{B_j}$ or $I_{\beta_j}^{A_j} = I_{\beta_j}^{B_j} = 1$ and $J_{\beta_j}^{A_j} \neq J_{\beta_j}^{B_j}$. Therefore, letting $\mathcal{J} = \{j : I_{\beta_j}^{A_j} \neq I_{\beta_j}^{B_j}\}$ and $\bar{\mathcal{J}} = \{1, \dots, r\} \setminus \mathcal{J}$, we obtain

$$\tilde{p} \leq \mathbb{P} \left[\bigwedge_{i=1}^l I_{\alpha_i} = 1 \wedge \bigwedge_{j \in \mathcal{J}} I_{\beta_j}^{A_j} \neq I_{\beta_j}^{B_j} \wedge \bigwedge_{j \in \bar{\mathcal{J}}} \left(I_{\beta_j}^{A_j} = 1 \wedge J_{\beta_j}^{A_j} \neq J_{\beta_j}^{B_j} \right) \right]. \quad (3.48)$$

Now, the random variables I_{α_i} , $I_{\beta_j}^{A_j}$, and $I_{\beta_j}^{B_j}$ are determined just by the edges in $\mathcal{E} \setminus \mathcal{E}(G)$, while $J_{\beta_j}^{A_j}$ and $J_{\beta_j}^{B_j}$ depend only on the edges in $\mathcal{E}(G)$. Hence, as the edges in $\mathcal{E} \setminus \mathcal{E}(G)$ and in $\mathcal{E}(G)$ occur in H independently, (3.48) yields

$$\tilde{p} \leq \mathbb{P} \left[\bigwedge_{i=1}^l I_{\alpha_i} = 1 \wedge \bigwedge_{j \in \bar{\mathcal{J}}} I_{\beta_j}^{A_j} = 1 \wedge \bigwedge_{j \in \mathcal{J}} I_{\beta_j}^{A_j} \neq I_{\beta_j}^{B_j} \right] \cdot \mathbb{P} \left[\bigwedge_{j \in \bar{\mathcal{J}}} J_{\beta_j}^{A_j} \neq J_{\beta_j}^{B_j} \right]. \quad (3.49)$$

Furthermore, Lemma 2.6 entails that

$$\begin{aligned} \mathbb{P} \left[\bigwedge_{i=1}^l I_{\alpha_i} = 1 \wedge \bigwedge_{j \in \bar{\mathcal{J}}} I_{\beta_j}^{A_j} = 1 \wedge \bigwedge_{j \in \mathcal{J}} I_{\beta_j}^{A_j} \neq I_{\beta_j}^{B_j} \right] \\ \leq O \left(n^{-|\mathcal{J}|} \cdot \text{polylog } n \right) \cdot \prod_{i=1}^l \mathbb{P} [I_{\alpha_i} = 1] \cdot \prod_{j=1}^r \mathbb{P} [I_{\beta_j} = 1]. \end{aligned} \quad (3.50)$$

In addition, we shall prove below that

$$\mathbb{P} \left[\bigwedge_{j \in \bar{\mathcal{J}}} J_{\beta_j}^{A_j} \neq J_{\beta_j}^{B_j} \right] \leq O \left(n^{-|\bar{\mathcal{J}}|} \cdot \text{polylog } n \right). \quad (3.51)$$

Plugging (3.50) and (3.51) into (3.49), we get $\tilde{p} \leq O(n^{-r} \cdot \text{polylog } n) \cdot \prod_{i=1}^l \mathbb{P} [I_{\alpha_i} = 1] \cdot \prod_{j=1}^r \mathbb{P} [I_{\beta_j} = 1]$, so that the assertion follows from (3.47).

Thus, the remaining task is to establish (3.51). Let us first deal with the case $|\bar{\mathcal{J}}| = 1$. Let $j \in \bar{\mathcal{J}}$. If $J_{\beta_j}^{A_j} \neq J_{\beta_j}^{B_j}$, then $J_{\beta_j}^{A_j} = 1$ and $J_{\beta_j}^{B_j} = 0$, because $A_j \subset B_j$. Thus, β_j is connected to G via an edge that is incident with $A_j \setminus B_j$; that is, $H \cap \mathcal{E}(\beta_j, B_j \setminus A_j) \neq \emptyset$. Since there are $|\mathcal{E}(\beta_j, B_j \setminus A_j)| \leq |\beta_j| \cdot |B_j| \cdot n^{d-2} = O(n^{d-2} \cdot \text{polylog } n)$ such edges to choose from, and because each such edge is present with probability $p_2 = O(n^{1-d})$, we conclude that $\mathbb{P} [J_{\beta_j}^{A_j} \neq J_{\beta_j}^{B_j}] \leq \mathbb{P} [H \cap \mathcal{E}(\beta_j, B_j \setminus A_j) \neq \emptyset] \leq O(n^{d-2} \cdot \text{polylog } n) p_2 = O(n^{-1} \cdot \text{polylog } n)$, whence we obtain (3.51).

Finally, suppose that $|\bar{\mathcal{J}}| = 2$. If $J_{\beta_j}^{A_j} \neq J_{\beta_j}^{B_j}$ for $j = 1, 2$, then there occur edges $e_j \in H \cap \mathcal{E}(\beta_j, B_j \setminus A_j)$ ($j = 1, 2$).

- 1st case:** $e_1 = e_2$. In this case $e_1 = e_2$ is incident with all four sets $\beta_j, B_j \setminus A_j$ ($j = 1, 2$). Hence, as the number of such edges is $\leq n^{d-4} \prod_{j=1}^2 |\beta_j| \cdot |B_j \setminus A_j| \leq O(n^{d-4} \cdot \text{polylog } n)$ and each such edge occurs with probability $p_2 = O(n^{1-d})$, the probability that the 1st case occurs is $O(n^{d-4} \cdot \text{polylog } n) p_2 = O(n^{-3} \cdot \text{polylog } n)$.
- 2nd case:** $e_1 \neq e_2$. There are $\leq |\beta_j| \cdot |B_j \setminus A_j| \cdot n^{d-2} \leq O(n^{d-2} \cdot \text{polylog } n)$ ways to choose e_j for $j = 1, 2$, each of which is present with probability $p_2 = O(n^{1-d})$ independently. Hence, the probability that the second case occurs is bounded by $\left(O(n^{d-2} \cdot \text{polylog } n) p_2\right)^2 \leq O(n^{-2} \cdot \text{polylog } n)$.

Thus, the bound (3.51) holds in both cases. □

Chapter 4

Bivariate Limit Theorems

Having proven the local limit theorem for $\mathcal{N}(H_d(n, p))$, we will now turn to the joint distribution of $\mathcal{N}, \mathcal{M}(H_d(n, p))$. This can be actually derived from the former result by exploiting the fact that the number of edges outside the giant component is an independent random variable, once we know the order of the giant component.

4.1 Results

Our first result is the *local limit theorem* for the joint distribution of $\mathcal{N}(H_d(n, p))$ and $\mathcal{M}(H_d(n, p))$.

Theorem 4.1. *Let $d \geq 2$ be a fixed integer. For any two compact sets $\mathcal{I} \subset \mathbb{R}^2$, $\mathcal{J} \subset ((d-1)^{-1}, \infty)$, and for any $\delta > 0$ there exists $n_0 > 0$ such that the following holds. Let $p = p(n)$ be a sequence such that $c = c(n) = \binom{n-1}{d-1}p \in \mathcal{J}$ for all n and let $0 < \rho = \rho(n) < 1$ be the unique solution to (1.2). Further, let*

$$\sigma_{\mathcal{N}}^2 = \frac{\rho \left(1 - \rho + c(d-1)(\rho - \rho^{d-1})\right)}{(1 - c(d-1)\rho^{d-1})^2} n, \quad (4.1)$$

$$\sigma_{\mathcal{M}}^2 = c^2 \rho^d \frac{2 + c(d-1)\rho^{2d-2} - 2c(d-1)\rho^{d-1} + c(d-1)\rho^d - \rho^{d-1} - \rho^d}{(1 - c(d-1)\rho^{d-1})^2} n + (1 - \rho^d) \frac{cn}{d},$$

$$\sigma_{\mathcal{NM}} = c\rho \frac{1 - \rho^d - c(d-1)\rho^{d-1}(1 - \rho)}{(1 - c(d-1)\rho^{d-1})^2} n.$$

If $n \geq n_0$ and if ν, μ are integers and $x := \nu - (1 - \rho)n$ and $y := \mu - (1 - \rho^d) \binom{n}{d} p$ are such that $n^{-\frac{1}{2}} \binom{x}{y} \in \mathcal{I}$, then letting

$$P(x, y) = \frac{1}{2\pi \sqrt{\sigma_{\mathcal{N}}^2 \sigma_{\mathcal{M}}^2 - \sigma_{\mathcal{NM}}^2}} \exp \left(-\frac{\sigma_{\mathcal{N}}^2 \sigma_{\mathcal{M}}^2}{2(\sigma_{\mathcal{N}}^2 \sigma_{\mathcal{M}}^2 - \sigma_{\mathcal{NM}}^2)} \left(\frac{x^2}{\sigma_{\mathcal{N}}^2} - 2\sigma_{\mathcal{NM}}^2 \frac{xy}{\sigma_{\mathcal{N}}^2 \sigma_{\mathcal{M}}^2} + \frac{y^2}{\sigma_{\mathcal{M}}^2} \right) \right) \quad (4.2)$$

we have

$$(1 - \delta)P(x, y) \leq \mathbb{P}[\mathcal{N}(H_d(n, m)) = \nu \wedge \mathcal{M}(H_d(n, m)) = \mu] \leq (1 + \delta)P(x, y).$$

Theorem 4.1 characterises the joint limiting distribution of $\mathcal{N}, \mathcal{M}(H_d(n, p))$ *precisely*, because it actually yields the asymptotic probability that \mathcal{N} and \mathcal{M} attain any two values $\nu = (1 - \rho)n + O(\sigma_{\mathcal{N}})$, $\mu = (1 - \rho^d)\binom{n}{d}p + O(\sigma_{\mathcal{M}})$ and it guarantees some uniformity of convergence. We emphasise that the expression on the r.h.s. of (4.2) is as small as $O(n^{-1})$ as $n \rightarrow \infty$.

Observe that the r.h.s. of (4.2) is just the density function of a bivariate normal distribution. Therefore, Theorem 4.1 readily yields the following *central limit theorem* for the joint distribution of $\mathcal{N}, \mathcal{M}(H_d(n, p))$.

Corollary 4.2. *Keep the notation from Theorem 4.1. Then $\sigma_{\mathcal{N}}^{-1}(\mathcal{N}(H_d(n, p)) - (1 - \rho)n), \sigma_{\mathcal{M}}^{-1}(\mathcal{M}(H_d(n, p)) - (1 - \rho^d)\binom{n}{d}p)$ converge to the bivariate normal distribution with mean 0 and covariance matrix $\begin{pmatrix} 1 & r \\ r & 1 \end{pmatrix}$ where $r = \frac{\sigma_{\mathcal{N}\mathcal{M}}}{\sigma_{\mathcal{N}}\sigma_{\mathcal{M}}}$.*

Nonetheless, we stress that Theorem 4.1 is considerably more precise than Corollary 4.2. The latter result just yields the asymptotic probability that $\sigma_{\mathcal{N}}^{-1}(\mathcal{N}(H_d(n, p)) - (1 - \rho)n) \in (a, a')$ and simultaneously $\sigma_{\mathcal{M}}^{-1}(\mathcal{M}(H_d(n, p)) - (1 - \rho^d)\binom{n}{d}p) \in (b, b')$ for any fixed $a, a', b, b' \in \mathbb{R}$. Hence, Corollary 4.2 just determines $\mathcal{N}, \mathcal{M}(H_d(n, p))$ up to errors of $o(\sigma_{\mathcal{N}})$ and $o(\sigma_{\mathcal{M}})$, while Theorem 4.1 actually yields the probabilities of hitting *exactly* specific values ν, μ .

The second main result of this paper is a local limit theorem for the joint distribution of $\mathcal{N}(H_d(n, m))$ and $\mathcal{M}(H_d(n, m))$.

Theorem 4.3. *Let $d \geq 2$ be a fixed integer. For any two compact sets $\mathcal{I} \subset \mathbb{R}^2$, $\mathcal{J} \subset ((d - 1)^{-1}, \infty)$, and for any $\delta > 0$ there exists $n_0 > 0$ such that the following holds. Let $m = m(n)$ be a sequence of integers such that $c = c(n) = dm/n \in \mathcal{J}$ for all n and let $0 < \rho = \rho(n) < 1$ be the unique solution to (1.2). Further, let*

$$\begin{aligned} \tilde{\sigma}_{\mathcal{N}}^2 &= \rho \frac{1 - (c + 1)\rho - c(d - 1)\rho^{d-1} + 2cd\rho^d - cd\rho^{2d-1}}{(1 - c(d - 1)\rho^{d-1})^2} n, \\ \tilde{\sigma}_{\mathcal{M}}^2 &= c\rho^d \frac{1 - c(d - 2)\rho^{d-1} - (c^2d - cd + 1)\rho^d - c^2(d - 1)\rho^{2d-2} + 2c(cd - 1)\rho^{2d-1} - c^2\rho^{3d-2}}{d(1 - c(d - 1)\rho^{d-1})^2} n, \\ \tilde{\sigma}_{\mathcal{N}\mathcal{M}} &= c\rho^d \frac{1 - c\rho - c(d - 1)\rho^{d-1} + (c + cd - 1)\rho^d - c\rho^{2d-1}}{(1 - c(d - 1)\rho^{d-1})^2} n. \end{aligned}$$

If $n \geq n_0$ and if ν, μ are integers and $x := \nu - (1 - \rho)n$ and $y := \mu - (1 - \rho^d)m$ are such that $n^{-\frac{1}{2}}\begin{pmatrix} x \\ y \end{pmatrix} \in \mathcal{I}$, then letting

$$P(x, y) = \frac{1}{2\pi\sqrt{\tilde{\sigma}_{\mathcal{N}}^2\tilde{\sigma}_{\mathcal{M}}^2 - \tilde{\sigma}_{\mathcal{N}\mathcal{M}}^2}} \exp\left(-\frac{\tilde{\sigma}_{\mathcal{N}}^2\tilde{\sigma}_{\mathcal{M}}^2}{2(\tilde{\sigma}_{\mathcal{N}}^2\tilde{\sigma}_{\mathcal{M}}^2 - \tilde{\sigma}_{\mathcal{N}\mathcal{M}}^2)} \left(\frac{x^2}{\tilde{\sigma}_{\mathcal{N}}^2} - 2\tilde{\sigma}_{\mathcal{N}\mathcal{M}}^2\frac{xy}{\tilde{\sigma}_{\mathcal{N}}^2\tilde{\sigma}_{\mathcal{M}}^2} + \frac{y^2}{\tilde{\sigma}_{\mathcal{M}}^2}\right)\right)$$

we have

$$(1 - \delta)P(x, y) \leq \mathbb{P}[\mathcal{N}(H_d(n, m)) = \nu \wedge \mathcal{M}(H_d(n, m)) = \mu] \leq (1 + \delta)P(x, y).$$

Similarly as Theorem 4.1, Theorem 4.3 characterises the joint limiting distribution of $\mathcal{N}, \mathcal{M}(H_d(n, m))$ precisely. Once more the limit resembles a bivariate normal distribution, so that we can infer the following central limit theorem.

Corollary 4.4. *Keep the notation from Theorem 4.3. Then $\tilde{\sigma}_{\mathcal{N}}^{-1}(\mathcal{N}(H_d(n, m)) - (1 - \rho)n), \tilde{\sigma}_{\mathcal{M}}^{-1}(\mathcal{M}(H_d(n, m)) - (1 - \rho^d)m)$ converge to the bivariate normal distribution with mean 0 and covariance matrix $\begin{pmatrix} 1 & \tilde{r} \\ \tilde{r} & 1 \end{pmatrix}$ where $\tilde{r} = \frac{\tilde{\sigma}_{\mathcal{N}\mathcal{M}}}{\tilde{\sigma}_{\mathcal{N}}\tilde{\sigma}_{\mathcal{M}}}$.*

The comparison of the correlation factors (cf. Figure 4.1) shows a much faster decay in correlation for growing c in the $H_d(n, m)$ model than in $H_d(n, p)$.

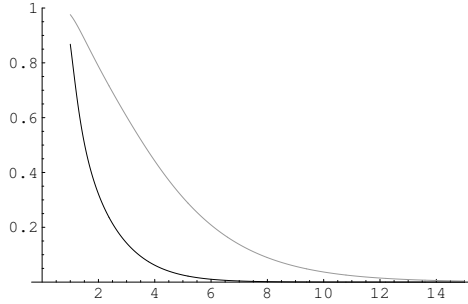


Figure 4.1: The correlation factors \tilde{r} (black) and r (gray) for 3-uniform hypergraphs in the range $1 < c < 15$.

4.2 Bivariate Limit Theorem

4.2.1 Outline

In order to prove Theorem 4.3, our starting point is Theorem 3.1, i.e., the local limit theorem for $\mathcal{N}(H_d(n, p))$; we shall convert this *univariate* limit theorem into a *bivariate* one that covers both \mathcal{N} and \mathcal{M} . To this end, we observe that Theorem 3.1 easily yields a local limit theorem for the joint distribution of $\mathcal{N}(H_d(n, p))$ and the number $\bar{\mathcal{M}}(H_d(n, p))$ of edges *outside* the largest component of $H_d(n, p)$. Indeed, it is easy to prove that *given* that $\mathcal{N}(H_d(n, p)) = \nu$ the random variable $\bar{\mathcal{M}}(H_d(n, p))$ has approximately a binomial distribution $\text{Bi}(\binom{n-\nu}{d}, p)$ (cf. Lemma 4.5 below). However, this does *not* yield a result on the joint distribution of $\mathcal{N}(H_d(n, p))$ and $\mathcal{M}(H_d(n, p))$. For the random variables $\mathcal{M}(H_d(n, p))$ and $\bar{\mathcal{M}}(H_d(n, p))$ are not directly related, because the *total* number of edges in $H_d(n, p)$ is a random variable.

Therefore, to derive the joint distribution of $\mathcal{N}(H_d(n, p))$ and $\mathcal{M}(H_d(n, p))$, we make a detour to the $H_d(n, m)$ model, in which the total number of edges is fixed (namely, m). Hence, in the $H_d(n, m)$ model the step from \mathcal{M} to $\bar{\mathcal{M}}$ is easy (because $\bar{\mathcal{M}}(H_d(n, m)) = m - \mathcal{M}(H_d(n, m))$). Moreover, $H_d(n, p)$ and $H_d(n, m)$ are related as follows: given that the total number of edges in $H_d(n, p)$ equals m , $H_d(n, p)$ is distributed as $H_d(n, m)$;

consequently,

$$\begin{aligned} \mathbb{P} \left[\mathcal{N}(H_d(n, p)) = \nu \wedge \bar{\mathcal{M}}(H_d(n, p)) = \bar{\mu} \right] \\ = \sum_{m=0}^{\binom{n}{d}} \mathbb{P} \left[\mathcal{N}(H_d(n, m)) = \nu \wedge \bar{\mathcal{M}}(H_d(n, m)) = \bar{\mu} \right] \cdot \mathbb{P} \left[\text{Bi} \left(\binom{n}{d}, p \right) = m \right]. \end{aligned} \quad (4.3)$$

Now we would like to “solve” (4.3) for $\mathbb{P} \left[\mathcal{N}(H_d(n, m)) = \nu \wedge \bar{\mathcal{M}}(H_d(n, m)) = \bar{\mu} \right]$. To this end, note that Theorem 3.1 yields an explicit expression for the l.h.s. of (4.3) (cf. Lemma 4.5), and that Proposition 1.1 provides an explicit formula for the second factor on the r.h.s. The crucial observation is that $\mathbb{P} \left[\mathcal{N}(H_d(n, m)) = \nu \wedge \bar{\mathcal{M}}(H_d(n, m)) = \bar{\mu} \right]$ is *independent of p* , while equation (4.3) is true *for all p* .

To exploit this observation, let $p_z = p + z\sigma \binom{n}{d}^{-1}$, where $\sigma^2 = \binom{n}{d}p(1-p)$. Moreover, let $m_z = \binom{n}{d}p + z\sigma$, set $z^* = \ln^2 n$, and consider the two functions

$$\begin{aligned} f(z) &= \begin{cases} n\mathbb{P} \left[\mathcal{N}(H_d(n, p_z)) = \nu \wedge \bar{\mathcal{M}}(H_d(n, p_z)) = \bar{\mu} \right] & \text{if } z \in (-z^*, z^*) \\ 0 & \text{if } z \in \mathbb{R} \setminus (-z^*, z^*), \end{cases} \\ g(z) &= \begin{cases} n\mathbb{P} \left[\mathcal{N}(H_d(n, m_z)) = \nu \wedge \bar{\mathcal{M}}(H_d(n, m_z)) = \bar{\mu} \right] & \text{if } z \in (-z^*, z^*) \\ 0 & \text{if } z \in \mathbb{R} \setminus (-z^*, z^*). \end{cases} \end{aligned} \quad (4.4)$$

Then computing the coefficients $\mathbb{P} \left[\mathcal{N}(H_d(n, m)) = \nu \wedge \bar{\mathcal{M}}(H_d(n, m)) = \bar{\mu} \right]$ is the same as computing the function g explicitly. To this end, we are going to show that (4.3) can be restated as $\|f - g * \phi\|_2 = o(1)$. Further, this relation in combination with some Fourier analysis will yield a formula for $g(z)$.

To see that (4.3) implies $\|f - g * \phi\|_2 = o(1)$, we first need to analyse the functions f and g a little. Using Theorem 3.1 and Proposition 1.1, we can estimate f as follows.

In the following we will analyse f and g in terms of x and y (instead of ν and $\bar{\mu}$) chosen such that $\nu = (1 - \rho)n + x$ and $\bar{\mu} = \rho^d m_0 - y = \rho^d \binom{n}{d}p - y$.

Lemma 4.5. *There exists a number $\gamma_0 > 0$ that remains fixed as $n \rightarrow \infty$ such that the following holds. For each $\gamma > \gamma_0$ there exists $n_0 > 0$ such that for all $n \geq n_0$ the following holds. Letting*

$$F(z) = \frac{n}{2\pi\sqrt{\rho^d\sigma\sigma_{\mathcal{N}}}} \exp \left(-\frac{(x - z\lambda_{\mathcal{N}})^2}{2\sigma_{\mathcal{N}}^2} - \frac{y - c\rho^{d-1}x + \rho^d\sigma z)^2}{2\rho^d\sigma^2} \right),$$

we have $|f(z) - F(z)| \leq \gamma^{-2}$ for all $z \in (-\gamma, \gamma)$. If $|z| > \gamma_0$, then $|f(z)| \leq \exp(-z^2/\gamma_0) + O(n^{-90})$.

We defer the proof of Lemma 4.5 to Section 4.2.3. Note that Lemma 4.5 provides an explicit expression $F(z)$ that approximates $f(z)$ well on compact sets, and shows that $f(z) \rightarrow 0$ rapidly as $z \rightarrow \infty$. Furthermore, the following lemma, whose proof we defer to Section 4.2.4, shows that g enjoys a certain “continuity” property.

Lemma 4.6. *For any $\alpha > 0$ there are $\beta > 0$ and $n_0 > 0$ such that for $n \geq n_0$ and $z, z' \in (-z^*, z^*)$ such that $|z - z'| < \beta$ we have $g(z') \leq (1 + \alpha)g(z) + n^{-20}$.*

Further, in Section 4.2.6 we shall combine Lemmas 4.5 and 4.6 to restate (4.3) as follows.

Lemma 4.7. *We have $f(z) = (1 + o(1))(g * \phi(z)) + O(n^{-18})$ for all $z \in \mathbb{R}$.*

Since f is bounded and both f and g vanish outside of the interval $(-z^*, z^*)$, Lemma 4.7 entails that $\|f - g * \phi\|_2 = o(1)$.

To obtain an explicit formula for g , we exhibit another function h such that $\|f - h * \phi\|_2 = o(1)$.

Lemma 4.8. *Let*

$$\chi := \left(\frac{d\sigma\rho(1 - \rho^{d-1})}{\sigma_{\mathcal{N}}(1 - c(d-1)\rho^{d-1})} \right)^2 + \rho^d \quad (4.5)$$

$$\kappa := -\frac{d\sigma\rho(1 - \rho^{d-1})}{\sigma_{\mathcal{N}}^2(1 - c(d-1)\rho^{d-1})}x - \frac{c\rho^{d-1}}{\sigma}x + \frac{1}{\sigma}y \quad (4.6)$$

$$\theta := \frac{1}{\sigma_{\mathcal{N}}^2}x^2 + \frac{(c\rho^{d-1}x - y)^2}{\rho^d\sigma^2} \quad (4.7)$$

as well as

$$h(z) = n \frac{\exp(-\frac{1}{2}(\theta - \frac{\kappa^2}{\chi}))}{\sigma_{\mathcal{N}}\sigma\rho^{d/2}2\pi\sqrt{1-\chi}} \exp\left(-\frac{\left(z + \frac{\kappa}{\chi}\right)^2}{2(\frac{1}{\chi} - 1)}\right) \quad (4.8)$$

Then $\|f - h * \phi\|_2 = o(1)$.

Proof. We know from Lemma 4.5 an explicit form of f , thus we just need to calculate the convolution of h with ϕ to get the desired result. The details of this merely technical calculation are deferred to Section 4.3. \square

Thus, we have the two relations $\|f - g * \phi\|_2 = o(1)$ and $\|f - h * \phi\|_2 = o(1)$. Using Fourier analysis, we shall prove in Section 4.2.2 that these bounds imply that actually h approximates g pointwise.

Lemma 4.9. *For any $\alpha > 0$ there is $n_0 > 0$ such that for all $n > n_0$ and all $z \in (-z^*/2, z^*/2)$ we have $|g(z) - h(z)| < \alpha$.*

In summary, we have obtained an explicit formula for $g(z)$ by rephrasing (4.3) in terms of f and g as $\|f - g * \phi\|_2 = o(1)$. Since Theorem 3.1 yields an explicit formula for f , we have been able to compute g from this relation via Fourier analysis. In particular, we have an asymptotic formula for $g(0) = \mathbb{P}[\mathcal{N}(H_d(n, m_0)) = \nu \wedge \bar{\mathcal{M}}(H_d(n, m_0)) = \bar{\mu}]$; which implies Theorem 4.3.

Proof of Theorem 4.3.

As already pointed out $g(0) = \mathbb{P}[\mathcal{N}(H_d(n, m_0)) = \nu \wedge \bar{\mathcal{M}}(H_d(n, m_0)) = \bar{\mu}]$ and $h(0) \sim g(0)$ by Lemma 4.9. Thus we only plug in the definitions of Lemma 4.8 to calculate $h(0)$ which is a merely technical computation and is deferred to Section 4.3. \square

Finally, let us derive Theorem 4.1 from Theorem 4.3.

Proof of Theorem 4.1. Since we will use $g(z)$ and $h(z)$ as defined by (4.8) and (4.4) for different values of x and y in the sequel we will denote them by $g(x, y, z)$ and $h(x, y, z)$. Similarly to the proof of Theorem 4.3, let $p_z = p + z\sigma \binom{n}{d}^{-1}$ and $m_z = m + z\sigma$ and define

$$\tilde{g}(x', y', z) := n\mathbb{P}[\mathcal{N}(H_d(n, m_z)) = \nu \wedge \mathcal{M}(H_d(n, m_z)) = \mu] \quad (4.9)$$

with $\nu = (1 - \rho)n + x'$ and $\mu = (1 - \rho^d)m + y'$. We know from Lemma 4.8 and Lemma 4.9 that

$$\begin{aligned} h(x, y, z) &= g(x, y, z) + o(1) \\ &= n\mathbb{P}[\mathcal{N}(H_d(n, m_z)) = (1 - \rho)n + x \wedge \bar{\mathcal{M}}(H_d(n, m_z)) = \rho^d m - y] + o(1) \\ &= n\mathbb{P}[\mathcal{N}(H_d(n, m_z)) = (1 - \rho)n + x \wedge \mathcal{M}(H_d(n, m_z)) = m_z - \rho^d m + y] + o(1) \end{aligned} \quad (4.10)$$

thus if we let $x' = x$ and $y' = y - z\sigma$ and plug in (4.10) we have

$$\tilde{g}(x', y', z) = h(x', y', z) + o(1) = h(x, y - z\sigma, z) + o(1).$$

Furthermore, due to a relationship similar to (4.3), we have

$$\begin{aligned} n\mathbb{P}[\mathcal{N}(H_d(n, p)) = \nu \wedge \mathcal{M}(H_d(n, p)) = \mu] &= n \sum_{m=0}^{\binom{n}{d}} \mathbb{P}[\mathcal{N}(H_d(n, m)) = \nu \wedge \mathcal{M}(H_d(n, m)) = \mu] \mathbb{P}[|E(H_d(n, p))| = m] \\ &= n \sum_{z\sigma = -m}^{\binom{n}{d} - m} \mathbb{P}[\mathcal{N}(H_d(n, m_z)) = \nu \wedge \mathcal{M}(H_d(n, m_z)) = \mu] \mathbb{P}\left[\text{Bi}\left(\binom{n}{d} p\right) = m_z\right] \\ &\sim \int_{-\frac{\ln^2 n}{\sigma}}^{\frac{\ln^2 n}{\sigma}} h(x, y - z\sigma, z) \phi(z) dz \sim \int_{-\infty}^{\infty} h(x, y - z\sigma, z) \phi(z) dz \end{aligned}$$

That we can limit our interest to the values of z such that $|z\sigma| < \ln^2 n$ is due to the fact that both h and ϕ get exponentially small (in n) for larger z . Furthermore the step from the sum to an integral is possible because the two functions are invariant to changes in z of order $\frac{1}{\sigma}$.

Calculating this convolution with the explicit terms given for h and ϕ proves the theorem and is deferred to Section 4.3. \square

4.2.2 Fourier Analysis

This section contains the proof of Proposition 4.9.

We define the Fourier transform as $\hat{\varphi}(y) = (2\pi)^{-\frac{1}{2}} \int \varphi(x) \exp(ixy) dx$. This ensures that

$$\|\varphi\|_2 = \|\hat{\varphi}\|_2 \quad [\text{provided that } \varphi \in L_1(\mathbb{R}) \cap L_2(\mathbb{R})]. \quad (4.11)$$

Lemma 4.10. *There is number $0 < K = O(1)$ such that $g(z) \leq Kf(z) + O(n^{-18})$ for all $z \in (-z^*, z^*)$.*

Proof. Let $z \in (-z^*, z^*)$. By Lemma 4.6 there is a number $\gamma > 0$ such that $g(z') \geq \frac{1}{2}g(z) - n^{-20}$ for all $z' \in (-z^*, z^*)$ that satisfy $|z - z'| \leq \gamma$. Therefore, Lemma 4.7 entails that

$$\begin{aligned} f(z) &= (1 + o(1)) \int g(z + \zeta) \phi(\zeta) d\zeta + O(n^{-18}) \\ &\geq \frac{g(z)}{2 + o(1)} \int_{(-z^*, z^*) \cap (z - \gamma, z + \gamma)} \phi(\zeta) d\zeta + O(n^{-18}) \geq \frac{\gamma g(z)}{10} + O(n^{-18}), \end{aligned}$$

whence the desired estimate follows. \square

Since the bounds on f obtained in Lemma 4.5 shows that $\|f\|_1$ and $\|f\|_2$ (exist and) remain bounded as $n \rightarrow \infty$, Lemma 4.10 implies that the same is true for g .

Corollary 4.11. *We have $\|g\|_1, \|g\|_2 = O(1)$ as $n \rightarrow \infty$.*

Thus, we can apply the Plancherel theorem (4.11) to both f and g .

Lemmas 4.7 and 4.8 imply that there is a function $\omega = \omega(n)$ such that $\lim_{n \rightarrow \infty} \omega(n) = \infty$ and $\|f - g * \phi\|_2, \|f - h * \phi\|_2 < \frac{1}{2} \exp(-\omega^2)$. Thus,

$$\|(g - h) * \phi\|_2 < \exp(-\omega^2) = o(1). \quad (4.12)$$

In order to compare g and h , the crucial step is to establish that actually $\|(g - h) * \phi_{0, \tau^2}\|_2 = o(1)$ for “small” numbers $\tau < 1$; indeed, we are mainly interested in $\tau = o(1)$.

Lemma 4.12. *Suppose that $1 \geq \tau \geq \omega^{-1/8}$. Then $\|(g - h) * \phi_{0, \tau^2}\|_2 \leq \exp(-\omega/5)$.*

Proof. Let $\xi = \hat{\phi}_{0, \tau^2} = \phi_{0, \tau^{-2}}$. Then

$$\|(g - h) * \phi_{0, \tau^2}\|_2^2 \stackrel{(4.11)}{=} \|(\hat{g} - \hat{h})\xi\|_2^2 = \int_{-\omega}^{\omega} |(\hat{g} - \hat{h})\xi|^2 + \int_{\mathbb{R} \setminus (-\omega, \omega)} |(\hat{g} - \hat{h})\xi|^2. \quad (4.13)$$

Since $\hat{\phi} = \phi$, we obtain

$$\begin{aligned} \int_{-\omega}^{\omega} |(\hat{g} - \hat{h})\xi|^2 &\leq \frac{\|\xi\|_{\infty}}{\inf_{-\omega \leq t \leq \omega} |\hat{\phi}(t)|^2} \int_{-\omega}^{\omega} |(\hat{g} - \hat{h})\hat{\phi}|^2 \\ &\leq \exp(\omega^2) \|(\hat{g} - \hat{h})\hat{\phi}\|_2^2 \stackrel{(4.11)}{=} \exp(\omega^2) \|(g - h) * \phi\|_2^2 \stackrel{(4.12)}{\leq} \exp(-\omega^2). \end{aligned} \quad (4.14)$$

In addition, by the Cauchy-Schwarz inequality

$$\int_{\mathbb{R} \setminus (-\omega, \omega)} |(\hat{g} - \hat{h})\xi|^2 \leq \left(\int_{\mathbb{R}} |(\hat{g} - \hat{h})^2|^2 \right)^{\frac{1}{2}} \cdot \left(\int_{\mathbb{R} \setminus (-\omega, \omega)} |\xi|^4 \right)^{\frac{1}{2}} \quad (4.15)$$

Furthermore, as $\tau^{-2} \leq \omega^{\frac{1}{4}}$, we have

$$\int_{\mathbb{R} \setminus (-\omega, \omega)} |\xi|^4 \leq \tau^{-2} \int_{\omega}^{\infty} \exp(-2\tau^2 \zeta^2) d\zeta \leq \exp(-\omega). \quad (4.16)$$

Moreover, by Lemma 4.10

$$\begin{aligned} \int_{\mathbb{R}} |(\hat{g} - \hat{h})^2|^2 &= \|(\hat{g} - \hat{h})^2\|_2^2 \stackrel{(4.11)}{=} \|(g - h) * (g - h)\|_2^2 \\ &\leq (\|g * g\|_2 + 2\|g * h\|_2 + \|h * h\|_2)^2 \end{aligned} \quad (4.17)$$

$$\leq \left(K^2 \|f * f\|_2 + 2K \|f * h\|_2 + \|h * h\|_2 \right)^2 + o(1). \quad (4.18)$$

Considering the bounds on f and h obtained in Lemma 4.5 and Lemma 4.8, we see that $\|f * f\|_2, \|f * h\|_2, \|h * h\|_2 = O(1)$. Therefore, (4.15), (4.16), and (4.18) imply that

$$\int_{\mathbb{R} \setminus (-\omega, \omega)} |(\hat{g} - \hat{h})\xi|^2 \leq O(\exp(-\omega/2)). \quad (4.19)$$

Finally, combining (4.13), (4.14), and (4.19), we obtain the desired bound on $\|(g - h) * \phi_{0, \tau^2}\|_2$. \square

Proof of Proposition 4.9. Assume for contradiction that there is some $z \in (-z^*/2, z^*/2)$ and some fixed $0 < \alpha = \Omega(1)$ such that $g(z) > h(z) + \alpha$ for arbitrarily large n (an analogous argument applies in the case $g(z) < h(z) - \alpha$). Let $\tau = \omega^{-1/8}$. Our goal is to show that in this case

$$\|(h - g) * \phi_{0, \tau^2}\|_2 > \exp(-\omega/5), \quad (4.20)$$

which contradicts Lemma 4.12.

To show (4.20), note that Lemma 4.10 implies that $\|g\|_{\infty} = O(1)$, because the bound $\|f\|_{\infty} = O(1)$ follows from Lemma 4.5. Similarly, the function h detailed in Lemma 4.8 is bounded. Thus, let $\kappa = O(1)$ be such that $g(\zeta), h(\zeta) \leq \kappa$ for all $\zeta \in \mathbb{R}$. Then Lemma 4.6 implies that there exists $0 < \beta = \Omega(1)$ such that

$$(1 - 0.01\alpha\kappa^{-1})g(z) - O(n^{-18}) \leq g(z') \leq (1 + 0.01\alpha\kappa^{-1})g(z) + O(n^{-18}) \text{ if } |z - z'| < \beta. \quad (4.21)$$

In fact, as h is continuous, we can choose β small enough so that in addition

$$(1 - 0.01\alpha\kappa^{-1})h(z) - O(n^{-18}) \leq h(z') \leq (1 + 0.01\alpha\kappa^{-1})h(z) + O(n^{-18}) \text{ if } |z - z'| < \beta. \quad (4.22)$$

Combining (4.21) and (4.22), we conclude that

$$|g(z') - g(z'')| \leq 0.1\alpha, \quad |h(z') - h(z'')| \leq 0.1\alpha \text{ for all } z', z'' \text{ such that } |z - z'|, |z - z''| < \beta. \quad (4.23)$$

Further, let $\gamma = \int_{\mathbb{R} \setminus (-\beta/2, \beta/2)} \phi_{0, \tau^2}$. Then for sufficiently large n we have $\gamma < 0.01\alpha\kappa^{-1}$, because $\tau \rightarrow 0$ as $n \rightarrow \infty$. Therefore, for any z' such that $|z' - z| < \beta/2$ we have

$$\begin{aligned} g * \phi_{0, \tau^2}(z') &= \int_{\mathbb{R}} g(z' + \zeta) \phi_{0, \tau^2}(\zeta) d\zeta \geq \int_{-\beta/2}^{\beta/2} g(z' + \zeta) \phi_{0, \tau^2}(\zeta) d\zeta \\ &\stackrel{(4.23)}{\geq} (g(z) - 0.01\alpha)(1 - \gamma) \geq g(z) - 0.02\alpha, \text{ and similarly} \end{aligned} \quad (4.24)$$

$$h * \phi_{0, \tau^2}(z') \leq h(z) + 0.02\alpha. \quad (4.25)$$

Since (4.24) and (4.25) are true for all z' such that $|z' - z| < \beta/2$, our assumption $g(z) > h(z) + \alpha$ yields

$$\|(g - h) * \phi_{0, \tau^2}\|_2^2 \geq \int_{-\beta/2}^{\beta/2} |g * \phi_{0, \tau^2}(z') - h * \phi_{0, \tau^2}(z')|^2 \geq 0.5\alpha^2\beta. \quad (4.26)$$

As α, β remain bounded away from 0 as $n \rightarrow \infty$, for sufficiently large n we have $0.5\alpha^2\beta > \exp(-\omega/5)$, so that (4.26) implies (4.20). \square

4.2.3 An Explicit Formula for the $H_d(n, p_z)$ Distribution $f(z)$

Lemma 4.13. *We have $\mathbb{E}[\mathcal{N}(H_d(n, p_z))] = \mu_{\mathcal{N}} + z\lambda_{\mathcal{N}} + o(\sqrt{n})$, where $\frac{\lambda_{\mathcal{N}}}{\sigma_{\mathcal{N}}} = \Theta(1)$ and $\mu_{\mathcal{N}} = \mathbb{E}[\mathcal{N}(H_d(n, p))]$.*

Proof. This follows from the fact that the function $c \mapsto \rho(c)$ is differentiable, which is an immediate consequence of the implicit function theorem. \square

Let $Q = \mathbb{P}[\mathcal{N}(H_d(n, p)) = \nu \wedge \bar{\mathcal{M}}(H_d(n, p)) = \bar{\mu}]$ and $N = \binom{n-\nu}{d}$. The crucial step in the proof of Lemma 4.5 is the derivation of the following estimate of Q .

Lemma 4.14. *We have $1 - n^{-98} \leq \frac{Q}{\mathbb{P}[\text{Bi}(N, p) = \bar{\mu}] \mathbb{P}[\mathcal{N}(H_d(n, p)) = \nu]} \leq 1 + n^{-98}$.*

Proof. Let $\mathcal{G} = \{G \subset V : |G| = \nu\}$. Moreover, for $G \in \mathcal{G}$ we let $\mathcal{C}_G(H)$ denote the event that G is a component in $H_d(n, p)$. Then by the union bound

$$\begin{aligned} Q &\leq \sum_{G \in \mathcal{G}} \mathbb{P}[\mathcal{C}_G \wedge |E(H_d(n, p) - G)| = \bar{\mu}] \\ &= \sum_{G \in \mathcal{G}} \mathbb{P}[\mathcal{C}_G] \mathbb{P}[|E(H_d(n, p) - G)| = \bar{\mu}]. \end{aligned} \quad (4.27)$$

As $H_d(n, p) - G$ is distributed as a random hypergraph $H_d(n - \nu, p)$, $|E(H_d(n, p) - G)|$ is binomially distributed with parameters N and p . Moreover, $H_d(n, p) - G$ is independent of G being a component. Therefore, (4.27) yields

$$\begin{aligned} Q &\leq \mathbb{P}[\text{Bi}(N, p) = \bar{\mu}] \sum_{G \in \mathcal{G}} \mathbb{P}[\mathcal{C}_G] \\ &= \mathbb{P}[\text{Bi}(N, p) = \bar{\mu}] \sum_{G \in \mathcal{G}} \frac{\mathbb{P}[\mathcal{C}_G \wedge \mathcal{N}(H_d(n, p) - G) < \nu]}{\mathbb{P}[\mathcal{N}(H_d(n, p) - G) < \nu]}. \end{aligned} \quad (4.28)$$

Furthermore, as we are assuming that $|\nu - (1 - \rho)n| = O(\sqrt{n})$, Theorem 1.2 implies that $\binom{n-\nu}{d-1}p < (d-1)^{-1}$. Consequently, $\mathbb{P}[\mathcal{N}(H_d(n, p) - G) < \nu] \geq 1 - n^{-100}$ (once more by Theorem 1.2). Thus, (4.28) entails that

$$\begin{aligned} (1 - n^{-100})\mathbb{P}[\text{Bi}(N, p) = \bar{\mu}]^{-1}Q &\leq \sum_{G \in \mathcal{G}} \mathbb{P}[\mathcal{C}_G \wedge \mathcal{N}(H_d(n, p) - G) < \nu] \\ &= \mathbb{P}[\exists G \in \mathcal{G} : \mathcal{C}_G \wedge \mathcal{N}(H_d(n, p) - G) < \nu] \\ &\leq \mathbb{P}[\mathcal{N}(H_d(n, p)) = \nu]. \end{aligned} \quad (4.29)$$

Conversely, if $G \in \mathcal{G}$ is a component of $H_d(n, p)$ and $\mathcal{N}(H_d(n, p) - G) < \nu$, then G is the unique largest component of $H_d(n, p)$. Therefore,

$$\begin{aligned} Q &\geq \sum_{G \in \mathcal{G}} \mathbb{P}[\mathcal{C}_G \wedge \mathcal{N}(H_d(n, p) - G) < \nu \wedge |E(H_d(n, p) - G)| = \bar{\mu}] \\ &= \sum_{G \in \mathcal{G}} \mathbb{P}[\mathcal{C}_G] \mathbb{P}[\mathcal{N}(H_d(n, p) - G) < \nu \wedge |E(H_d(n, p) - G)| = \bar{\mu}]. \end{aligned} \quad (4.30)$$

Further, given that $|E(H_d(n, p) - G)| = \bar{\mu}$, $H_d(n, p) - G$ is just a random hypergraph $H_d(n - \nu, \bar{\mu})$. Hence, (4.30) yields

$$\begin{aligned} Q &\geq \mathbb{P}[\mathcal{N}(H_d(n - \nu, \bar{\mu})) < \nu] \mathbb{P}[\text{Bi}(N, p) = \bar{\mu}] \sum_{G \in \mathcal{G}} \mathbb{P}[\mathcal{C}_G] \\ &\geq \mathbb{P}[\mathcal{N}(H_d(n - \nu, \bar{\mu})) < \nu] \mathbb{P}[\text{Bi}(N, p) = \bar{\mu}] \mathbb{P}[\mathcal{N}(H_d(n, p)) = \nu], \end{aligned} \quad (4.31)$$

where the last estimate follows from the union bound. Moreover, we claim that

$$\mathbb{P}[\mathcal{N}(H_d(n - \nu, \bar{\mu})) \geq \nu] \leq n^{-99}. \quad (4.32)$$

To see this, let $p' = \bar{\mu}/N$. Then by Proposition 1.1

$$n^{-1} \mathbb{P}[\mathcal{N}(H_d(n - \nu, \bar{\mu})) \geq \nu] \geq \mathbb{P}[\text{Bi}(N, p) = \bar{\mu}] \mathbb{P}[\mathcal{N}(H_d(n - \nu, \bar{\mu})) \geq \nu] \leq \mathbb{P}[\mathcal{N}(H_d(n - \nu, p') \geq \nu], \quad (4.33)$$

because given that $|E(H_d(n - \nu, p'))| = \bar{\mu}$, $H_d(n - \nu, p')$ has the same distribution as $H_d(n - \nu, \bar{\mu})$. Furthermore, as by assumption $\bar{\mu} \sim \binom{n-\nu}{d}p$, Theorem 1.2 entails that $\binom{n-\nu}{d-1}p' \sim \binom{n-\nu}{d-1}p < (d-1)^{-1}$. Hence, we obtain $\mathbb{P}[\mathcal{N}(H_d(n - \nu, p') \geq \nu] \leq n^{-100}$, so that (4.32) follows from (4.33). Thus, plugging (4.33) into (4.31), we get

$$Q \geq (1 - n^{-99})\mathbb{P}[\text{Bi}(N, p) = \bar{\mu}] \mathbb{P}[\mathcal{N}(H_d(n, p)) = \nu]. \quad (4.34)$$

Combining (4.29) and (4.34) completes the proof. \square

Proof of Lemma 4.5. To compare f and F , we fix some $\gamma > 0$, and consider $z \in (-\gamma, \gamma)$. Then Lemma 4.13 entails that $|\nu - \mathbb{E}[\mathcal{N}(H_d(n, p_z))]| = O(1)$ as $n \rightarrow \infty$. Therefore, Theorem 3.1 implies that

$$\begin{aligned} \mathbb{P}[\mathcal{N}(H_d(n, p_z)) = \nu] &\sim \frac{1}{\sqrt{2\pi\sigma_{\mathcal{N}}}} \exp\left(-\frac{(\nu - \mathbb{E}[\mathcal{N}(H_d(n, p_z))])^2}{2\sigma_{\mathcal{N}}^2}\right) \\ &\stackrel{\text{Lemma 4.13}}{\sim} \frac{1}{\sqrt{2\pi\sigma}} \exp\left(-\frac{(\nu - (1 - \rho)n - \lambda_{\mathcal{N}}z)^2}{2\sigma_{\mathcal{N}}^2}\right). \end{aligned} \quad (4.35)$$

Plugging (4.35) into the expression for Q from Lemma 4.14 and estimating $\mathbb{P}[\text{Bi}(N, p) = \bar{\mu}]$ via Proposition 1.1, we obtain the first part of Lemma 4.5. The detailed calculation can be found in Section 4.3.

To establish the second part, let us assume that $\gamma_0 < |z| \leq |z^*|$ for some large enough but fixed $\gamma_0 > 0$. Then $|Np_z - \bar{\mu}| = \Omega(z\sqrt{n})$. Therefore, Proposition 1.1 implies that $\mathbb{P}[\text{Bi}(n, p_z) = \bar{\mu}] \leq n^{-1/2} \exp(-\Omega(z^2))$. Furthermore, $\mathbb{P}[\mathcal{N}(H_d(n, p_z)) = \nu] = O(n^{-1/2})$ by Theorem 3.1. Hence, Lemma 4.14 entails that $Q \leq O(n^{-1} \exp(-\Omega(z^2)) + n^{-98})$, as desired. \square

4.2.4 Continuity of $g(z)$

Throughout this section we assume that $z, z' \in (-z^*, z^*)$, and that $|z - z'| < \beta$ for some small enough $\beta > 0$. In addition, we may assume that

$$g(z') \geq n^{-30}, \quad (4.36)$$

because otherwise the assertion is trivially true. To compare $g(z)$ and $g(z')$, we first express $g(z)$ in terms of the number $C(\nu, m_z - \bar{\mu})$ of connected d -uniform hypergraphs of order ν and size $m_z - \bar{\mu}$.

Lemma 4.15. *We have $\binom{n}{m} g(z) \sim n \binom{n}{\nu} C(\nu, m_z - \bar{\mu}) \binom{n-\nu}{\bar{\mu}} \binom{n}{m}^{-1}$. A similar statement is true for $g(z')$.*

Proof. We claim that

$$n^{-1} g(z) \leq \binom{n}{\nu} C(\nu, m_z - \bar{\mu}) \binom{n-\nu}{\bar{\mu}} \binom{n}{m}^{-1}. \quad (4.37)$$

The reason is that the right hand side equals the *expected* number of components of order ν and size $m_z - \bar{\mu}$ in $H_d(n, m)$. For there are $\binom{n}{\nu}$ ways to choose ν vertices where to place such a component. Then, there are $C(\nu, m_z - \bar{\mu})$ ways to choose the component itself. Moreover, there are $\binom{n-\nu}{\bar{\mu}}$ ways to choose the hypergraph induced on the remaining $n - \nu$ vertices, while the total number of d -uniform hypergraphs of order n and size m is $\binom{n}{m}$. Conversely,

$$n^{-1} g(z) \geq \binom{n}{\nu} C(\nu, m_z - \bar{\mu}) \binom{n-\nu}{\bar{\mu}} \mathbb{P}[\mathcal{N}(H_d(n - \nu, \mu)) < \nu] \binom{n}{m}^{-1}. \quad (4.38)$$

For the right hand side equals the probability that $H_d(n, m_z)$ has one component of order ν and size $m_z - \bar{\mu}$, while all other components have order $< \nu$.

Since $\mathbb{P}[\mathcal{N}(H_d(n - \nu, \mu)) < \nu] \sim 1$ by Theorem 1.2, the assertion follows from (4.37) and (4.38). \square

Lemma 4.15 entails that

$$\frac{g(z')}{g(z)} \sim \frac{C(\nu, m_{z'} - \mu)}{C(\nu, m_z - \mu)} \cdot \frac{\binom{n}{m_{z'}}}{\binom{n}{m_z}}. \quad (4.39)$$

Thus, as a next step we estimate the two factors on the r.h.s. of (4.39).

Lemma 4.16. *If $|z - z'| < \beta$ for a small enough $\beta > 0$, then $\frac{C(\nu, m_{z'} - \mu)}{C(\nu, m_z - \mu)} \cdot p^{m_z - m_{z'}} \leq 1 + \alpha/2$.*

To prove Lemma 4.16, we employ the following estimate, which we will establish in Section 4.2.5.

Lemma 4.17. *If $|z - z'| < \beta$ for a small enough $\beta > 0$, then letting*

$$\begin{aligned} P &= \mathbb{P}[\mathcal{N}(H_d(n, p_{z'})) = \nu \wedge \mathcal{N}(H_d(n, p_{z'})) = m_z - \mu], \\ P' &= \mathbb{P}[\mathcal{N}(H_d(n, p_{z'})) = \nu \wedge \mathcal{N}(H_d(n, p_{z'})) = m_{z'} - \mu], \end{aligned}$$

we have $(1 + \alpha/3)P - n^{-80} \geq P' \leq (1 + \alpha/3)P + n^{-80}$.

Proof of Lemma 4.16. We observe that

$$P \leq \binom{n}{\nu} C(\nu, m_z - \mu) p_{z'}^{m_z - \mu} (1 - p_{z'})^{\binom{n}{d} - \binom{n-\nu}{d} - (m_z - \mu)}, \quad (4.40)$$

because the r.h.s. equals the *expected* number of components of order ν and size $m_z - \mu$ in $H_d(n, p_{z'})$. (For there are $\binom{n}{\nu}$ ways to choose the ν vertices where to place the component and $C(\nu, m_z - \mu)$ ways to choose the component itself. Furthermore, edges are present with probability $p_{z'}$ independently, and thus the $p_{z'}^{m_z - \mu}$ factor accounts for the presence of the $m_z - \mu$ desired edges among the selected ν vertices, while the $(1 - p_{z'})$ -factor rules out further edges among the ν chosen vertices and inbetween the ν chosen and the $n - \nu$ remaining vertices.) Conversely,

$$P \geq \binom{n}{\nu} C(\nu, m_z - \mu) p_{z'}^{m_z - \mu} (1 - p_{z'})^{\binom{n}{d} - \binom{n-\nu}{d} - (m_z - \mu)} \mathbb{P}[\mathcal{N}(H_d(n - \nu, p_{z'})) < \nu]; \quad (4.41)$$

for the r.h.s. is the probability that there occurs exactly one component of order ν and size $m_z - \mu$, while all other components have order $< \nu$. As $p_{z'} \sim p$ and $n - \nu \sim (1 - \rho)n$, Theorem 1.2 entails that $\mathbb{P}[\mathcal{N}(H_d(n - \nu, p_{z'})) < \nu] \sim 1$. Consequently, (4.40) and (4.41) yield

$$P \sim \binom{n}{\nu} C(\nu, m_z - \mu) p_{z'}^{m_z - \mu} (1 - p_{z'})^{\binom{n}{d} - \binom{n-\nu}{d} - (m_z - \mu)}, \text{ and similarly} \quad (4.42)$$

$$P' \sim \binom{n}{\nu} C(\nu, m_{z'} - \mu) p_{z'}^{m_{z'} - \mu} (1 - p_{z'})^{\binom{n}{d} - \binom{n-\nu}{d} - (m_{z'} - \mu)}. \quad (4.43)$$

Therefore,

$$\begin{aligned} \frac{C(\nu, m_{z'} - \mu)}{C(\nu, m_z - \mu)} &\sim \frac{P'}{P} \cdot p_{z'}^{m_{z'} - m_z} \cdot (1 - p_{z'})^{m_z - m_{z'}} \sim \frac{P'}{P} \cdot p^{m_{z'} - m_z} \\ &\stackrel{\text{Lemma 4.17}}{\leq} \left(1 + \frac{\alpha}{3} + \frac{2}{n^{80} P' - 2}\right) p^{m_{z'} - m_z}. \end{aligned} \quad (4.44)$$

In order to show that the r.h.s. of (4.44) is $\leq 1 + \alpha/2$, we need to lower bound P' . Indeed, by Proposition 1.1

$$\begin{aligned} P' &\geq \mathbb{P}[\mathcal{N}(H_d(n, m_{z'})) = \nu \wedge \mathcal{M}(H_d(n, m_{z'})) = m_{z'} - \mu] \cdot \mathbb{P}\left[\text{Bi}\left(\binom{n}{d}, p_{z'}\right) = m_{z'}\right] \\ &\geq n^{-1} g(z') \stackrel{(4.36)}{\geq} n^{-31}. \end{aligned} \quad (4.45)$$

Finally, combining (4.44) and (4.45), we obtain the desired bound on $C(\nu, m_{z'} - \mu)$. \square

Lemma 4.18. *We have $\binom{n}{m_{z'}} \binom{n}{m_z}^{-1} = \exp(O(z - z')^2) \cdot p^{m_z - m_{z'}}$.*

Proof. By Stirling's formula,

$$\begin{aligned} \binom{n}{m_{z'}} \binom{n}{m_z}^{-1} &\sim \left(\frac{\binom{n}{d}}{m_{z'}}\right)^{m_{z'}} \left(\frac{\binom{n}{d}}{\binom{n}{d} - m_{z'}}\right)^{\binom{n}{d} - m_{z'}} \left(\left(\frac{\binom{n}{d}}{m_z}\right)^{m_z} \left(\frac{\binom{n}{d}}{\binom{n}{d} - m_z}\right)^{\binom{n}{d} - m_z}\right)^{-1} \\ &\sim \frac{p_z^{m_z}}{p_{z'}^{m_{z'}}} \left(1 + \frac{m_{z'}}{\binom{n}{d} - m_{z'}}\right)^{\binom{n}{d} - m_{z'}} \left(1 + \frac{m_z}{\binom{n}{d} - m_z}\right)^{m_z - \binom{n}{d}} \\ &\sim \frac{p_z^{m_z}}{p_{z'}^{m_{z'}}} \exp(m_{z'} - m_z) \sim p^{m_z - m_{z'}} \left(\frac{p_z}{p_{z'}}\right)^{m_{z'}} \exp(\sigma(z' - z)), \text{ where} \end{aligned} \quad (4.46)$$

$$\begin{aligned} \left(\frac{p_z}{p_{z'}}\right)^{m_{z'}} &\sim \left(\frac{m_0 + z\sigma}{m_0 + z'\sigma}\right)^{m_{z'}} \sim \exp\left((z - z')\sigma_0 - \frac{(z - z')^2\sigma_0^2}{2m_{z'}}\right) \\ &= \exp\left((z - z')\sigma_0 - O(z - z')^2\right) \end{aligned} \quad (4.47)$$

Combining (4.46) and (4.47), we obtain the assertion. \square

Plugging the estimates from Lemmas 4.16 and 4.18 into (4.39), we conclude that $1 - \alpha \leq g(z)/g(z') \leq 1 + \alpha$, provided that $|z - z'| < \beta$ for some small enough $\beta > 0$.

4.2.5 Proof of Lemma 4.16

By symmetry, it suffices to prove that $P' \leq (1 + \alpha/3)P + n^{-90}$. To show this, we expose the edges of $H_d(n, p_{z'})$ in three rounds. Let $\varepsilon > 0$ be a small enough number that remains fixed as $n \rightarrow \infty$. Moreover, set $q_1 = (1 - \varepsilon)p_{z'}$, and let $q_2 \sim \varepsilon p_{z'}$ be such that $q_1 + q_2 - q_1 q_2 = p_{z'}$. Choose ε sufficiently small, we can ensure that $\binom{n-1}{d-1} q_1 > (d-1)^{-1} + \varepsilon$. Now, we construct $H_d(n, p_{z'})$ in three rounds as follows.

1st round. Construct a random hypergraph H_1 with vertex set $V = \{1, \dots, n\}$ by including each of the $\binom{n}{d}$ possible edges with probability q_1 independently. Let G_1 be the largest component of H_1 .

2nd round. Let H_2 be the hypergraph obtained by adding with probability q_2 independently each possible edge $e \notin H_1$ that is not entirely contained in G_1 (i.e., $e \not\subset G_1$) to H_1 . Let G_2 signify the largest component of H_2 .

3rd round. Finally, obtain H_3 by adding each edge $e \notin H_1$ such that $e \subset G_1$ with probability q_2 independently. Let F denote the set of edges added in this way.

Since for each of the $\binom{n}{d}$ possible edges the overall probability of being contained in H_3 is $q_1 + (1 - q_1)q_2 = p_{z'}$, H_3 is just a random hypergraph $H_d(n, p_{z'})$. Moreover, as in the 3rd round we only add edges that fall completely into the component of H_2 that contains G_1 , we have $\mathcal{N}(H_d(n, p_{z'})) = \mathcal{N}(H_3) = \mathcal{N}(H_2)$. Furthermore, $|F|$ has a binomial distribution

$$|F| = \text{Bi}\left(\binom{|G_1|}{d} - \mathcal{M}(H_1), p_2\right). \quad (4.48)$$

To compare P' and P , we make use of the local limit theorem for the binomially distributed $|F|$ (Proposition 1.1): loosely speaking, we shall observe that most likely G_1 is contained in the largest component of H_3 . If this is indeed the case, then $\mathcal{M}(H_3) = |F| + \mathcal{M}(H_2)$, so that

$$\mathcal{M}(H_3) = m_{z'} - \mu \Leftrightarrow |F| = m_{z'} - \mu - \mathcal{M}(H_2), \quad (4.49)$$

$$\mathcal{M}(H_3) = m_z - \mu \Leftrightarrow |F| = m_z - \mu - \mathcal{M}(H_2). \quad (4.50)$$

Finally, since $\mathbb{P}[|F| = m_{z'} - \mu - \mathcal{M}(H_2)]$ is “close” to $\mathbb{P}[|F| = m_z - \mu - \mathcal{M}(H_2)]$ if $|z - z'|$ is small (by the local limit theorem), we conclude that P' cannot exceed P “significantly”.

To implement the above sketch, let \mathcal{Q} be the set of all pairs $(\mathcal{H}_1, \mathcal{H}_2)$ of hypergraphs that satisfy the following three conditions.

Q1. $\mathcal{N}(\mathcal{H}_2) = \nu$.

Q2. $\mathbb{P}[\mathcal{M}(H_3) = m_{z'} - \mu | H_1 = \mathcal{H}_1, H_2 = \mathcal{H}_2] \geq n^{-100}$.

Q3. The largest component of \mathcal{H}_2 contains the largest component of \mathcal{H}_1 .

The next lemma shows that the processes such that $(H_1, H_2) \in \mathcal{Q}$ constitute the dominant contribution.

Lemma 4.19. *Letting $P'' = \mathbb{P}[\mathcal{M}(H_3) = m_{z'} - \mu \wedge (H_1, H_2) \in \mathcal{Q}]$, we have $P' \leq P'' + n^{-99}$.*

Proof. Let \mathcal{R} signify the set of all pairs $(\mathcal{H}_1, \mathcal{H}_2)$ such that **Q1** is satisfied. Since $H_3 = H_d(n, p_{z'})$, we have $P' = \mathbb{P}[\mathcal{M}(H_3) = m_{z'} - \mu \wedge (H_1, H_2) \in \mathcal{R}]$. Therefore, letting $\bar{\mathcal{Q}}_2$ (resp. $\bar{\mathcal{Q}}_3$) denote the set of all $(\mathcal{H}_1, \mathcal{H}_2) \in \mathcal{R}$ that violate **Q2** (resp. **Q3**), we have

$$\begin{aligned} P' - P'' &\leq \mathbb{P}[\mathcal{M}(H_3) = m_{z'} - \mu \wedge (H_1, H_2) \in \mathcal{R} \setminus \mathcal{Q}] \\ &\leq \mathbb{P}[\mathcal{M}(H_3) = m_{z'} - \mu | (H_1, H_2) \in \bar{\mathcal{Q}}_2] + \mathbb{P}[(H_1, H_2) \in \bar{\mathcal{Q}}_3] \\ &\stackrel{\text{Q2}}{\leq} n^{-100} + \mathbb{P}[(H_1, H_2) \in \bar{\mathcal{Q}}_3]. \end{aligned} \quad (4.51)$$

Furthermore, if $(H_1, H_2) \in \bar{\mathcal{Q}}_3$, then either H_1 does not feature a component of order $\Omega(n)$, or H_2 has two such components. Since $\binom{n-1}{d-1}q_1 > (d-1)^{-1} + \varepsilon$ due to our choice of $\varepsilon > 0$, Theorem 1.2 entails that the probability of either event is $\leq n^{-100}$. Thus, the assertion follows from (4.51). \square

Finally, we can compare P and P'' as follows.

Lemma 4.20. *We have $P'' \leq (1 + \alpha/3)P$.*

Proof. Consider $(\mathcal{H}_1, \mathcal{H}_2) \in \mathcal{Q}$ and let us condition on the event $(H_1, H_2) = (\mathcal{H}_1, \mathcal{H}_2)$. Let $\Delta = m_z - \mu - \mathcal{M}(H_2)$, $\Delta' = m'_z - \mu - \mathcal{M}(H_2)$. We claim that

$$\left| \left(\binom{\nu}{d} - \mathcal{M}(H_1) \right) p_2 - \Delta' \right| \leq n^{0.51}; \quad (4.52)$$

for if $\left| \left(\binom{\nu}{d} - \mathcal{M}(H_1) \right) p_2 - \Delta' \right| > n^{0.51}$, then the Chernoff bound (1.3) entails that

$$\begin{aligned} \mathbb{P}[\mathcal{M}(H_3) = m_{z'} - \mu | (H_1, H_2) = (\mathcal{H}_1, \mathcal{H}_2)] &\stackrel{(4.49)}{=} \mathbb{P}[|F| = \Delta' | (H_1, H_2) = (\mathcal{H}_1, \mathcal{H}_2)] \\ &\stackrel{(4.48)}{\leq} \exp(-n^{0.01}) < n^{-100}, \end{aligned}$$

in contradiction to **Q2**. Thus, if $|z - z'| < \beta$ for a small enough $\beta > 0$, then Proposition 1.1 yields

$$\mathbb{P}[|F| = \Delta' | (H_1, H_2) = (\mathcal{H}_1, \mathcal{H}_2)] \leq (1 + \alpha/3) \mathbb{P}[|F| = \Delta | (H_1, H_2) = (\mathcal{H}_1, \mathcal{H}_2)], \quad (4.53)$$

because $|\Delta' - \Delta| = |z' - z|\sigma$, and $\text{Var}[|F|] \sim \binom{\nu}{d} p_2 = \Omega(\sigma^2)$. Since (4.53) holds for all $(\mathcal{H}_1, \mathcal{H}_2) \in \mathcal{Q}$, the assertion follows. \square

Finally, Lemma 4.16 is an immediate consequence of Lemmas 4.19 and 4.20.

4.2.6 Convolution

Now we prove Lemma 4.7. Set $m_- = m_0 - z^* \sigma$, $m_+ = m_0 + z^* \sigma$, and let

$$P(m) = n \mathbb{P}[\mathcal{N}(H_d(n, m)) = \nu \wedge \bar{\mathcal{M}}(H_d(n, m)) = \bar{\mu}], \quad B_z(m) = \mathbb{P}\left[\text{Bi}\left(\binom{n}{d}, p_z\right) = m\right].$$

Then for all $z \in (-z^*, z^*)$ we have

$$\begin{aligned} f(z) &= \sum_{m=0}^{\binom{n}{d}} P(m) B_z(m) \\ &\leq n \cdot \mathbb{P}\left[\text{Bi}\left(\binom{n}{d}, p_z\right) \notin (m_-, m_+)\right] + \sum_{m_- \leq m \leq m_+} P(m) B_z(m) \\ &\stackrel{(1.3)}{\leq} n^{-100} + \sum_{m_- \leq m \leq m_+} P(m) B_z(m). \end{aligned}$$

because $0 \leq P(m) \leq n$. Hence,

$$f(z) = O(n^{-100}) + \sum_{m_- \leq m \leq m_+} P(m)B_z(m). \quad (4.54)$$

We decompose the interval $J = (m_-, m_+)$ into k subsequent pieces J_1, \dots, J_k of lengths inbetween $\frac{\sigma}{2 \log n}$ and $\frac{\sigma}{\log n}$. Then Lemma 4.6 entails that

$$P(m) = (1 + o(1))P(m') + O(n^{-20}) \quad \text{for all } m, m' \in J_i \text{ and all } 1 \leq i \leq k. \quad (4.55)$$

Moreover, Proposition 1.1 yields that

$$B_z(m) \sim \frac{1}{\sqrt{2\pi}\sigma_0} \exp\left(-\frac{(m - m_z)^2}{2\sigma_0^2}\right) \quad \text{for all } m, m' \in J_i \text{ and all } 1 \leq i \leq k. \quad (4.56)$$

Further, let $I_i = \{\sigma^{-1}(x - m_0) : x \in J_i\}$ and set $M_i = \min J_i \cap \mathbb{Z}$. Combining (4.55) and (4.56), we obtain

$$\begin{aligned} \sum_{m \in J_i} P(m)B_z(m) &= O(n^{-18}) + (1 + o(1))P(M_i) \sum_{m \in J_i} B_z(m) \\ &= (1 + o(1))P(M_i) \int_{I_i} \phi(\zeta - z) d\zeta + O(n^{-18}) \\ &= (1 + o(1)) \int_{I_i} P(m_\zeta) \phi(\zeta - z) d\zeta + O(n^{-18}). \end{aligned} \quad (4.57)$$

As $|\zeta| \leq z^*$ for all $\zeta \in I_i$, we have $P(m_\zeta) = g(\zeta)$. Therefore, (4.57) yields

$$\sum_{m \in J_i} P(m)B_z(m) = (1 + o(1)) \int_{I_i} g(\zeta) \phi(\zeta - z) d\zeta + O(n^{-18}). \quad (4.58)$$

Summing (4.58) for $i = 1, \dots, k$, we get

$$f(z) \stackrel{(4.54)}{=} O(n^{-18}) + (1 + o(1)) \sum_{i=1}^k \int_{I_i} g(\zeta) \phi(\zeta - z) d\zeta \quad (4.59)$$

$$= O(n^{-18}) + (1 + o(1)) \int_{-z^*}^{z^*} g(\zeta) \phi(\zeta - z) d\zeta. \quad (4.60)$$

As $f(\zeta) = g(\zeta) = 0$ if $|\zeta| > z^*$, the assertion follows from (4.60).

4.3 Calculations

In order to keep the merely technical calculations apart from the core argument we have deferred some arguments to this section. This is probably the most technical part of the whole thesis and most of the calculations could be done using a computer algebra system and are presented for the sake of completeness only.

Let us first recall the most important definitions.

$$p_0 := c / \binom{n-1}{d-1} \quad (4.61)$$

$$m_0 := \binom{n}{d} p_0 = \frac{cn}{d} \quad (4.62)$$

$$\sigma^2 := \binom{n}{d} p_0 (1 - p_0) \sim \frac{cn}{d} \quad (4.63)$$

$$p_z := p_0 + z\sigma / \binom{n}{d} \quad (4.64)$$

$$m_z := \binom{n}{d} p_z = m_0 + z\sigma \quad (4.65)$$

$$\sigma_{\mathcal{N}}^2 := \frac{\rho \left(1 - \rho + c(d-1)(\rho - \rho^{d-1})\right)}{(1 - c(d-1)\rho^{d-1})^2} n \quad (4.66)$$

and furthermore

$$\lambda_{\mathcal{N}} := \frac{d\sigma\rho(1 - \rho^{d-1})}{1 - c(d-1)\rho^{d-1}} \quad (4.67)$$

$$\chi := \left(\frac{d\sigma\rho(1 - \rho^{d-1})}{\sigma_{\mathcal{N}}(1 - c(d-1)\rho^{d-1})} \right)^2 + \rho^d = \left(\frac{\lambda_{\mathcal{N}}}{\sigma_{\mathcal{N}}} \right)^2 + \rho^d \quad (4.68)$$

$$\kappa := -\frac{d\sigma\rho(1 - \rho^{d-1})}{\sigma_{\mathcal{N}}^2(1 - c(d-1)\rho^{d-1})} x - \frac{c\rho^{d-1}}{\sigma} x + \frac{1}{\sigma} y = -\frac{\lambda_{\mathcal{N}}}{\sigma_{\mathcal{N}}^2} x - \frac{c\rho^{d-1}}{\sigma} x + \frac{1}{\sigma} y \quad (4.69)$$

$$\theta := \frac{1}{\sigma_{\mathcal{N}}^2} x^2 + \frac{(c\rho^{d-1}x - y)^2}{\rho^d \sigma^2} \quad (4.70)$$

We do convolutions with the distribution of the number of edges in $H_d(n, p)$

$$\mathbb{P}[|E(H_d(n, p))| = m_z] = \mathbb{P}\left[\frac{|E(H_d(n, p))| - m_0}{\sigma} = z\right] \sim \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{z^2}{2}\right) =: \phi(z). \quad (4.71)$$

4.3.1 The Distribution for $H_d(n, m)$

We are interested in

$$\mathbb{P}\left[\mathcal{N}(H_d(n, m_z)) = (1 - \rho)n + x \wedge \mathcal{M}(H_d(n, m_z)) = (1 - \rho^d)m_0 + y\right]. \quad (4.72)$$

We first calculate the expectation of $\mathcal{N}(H_d(n, p_z))$. Let ρ_z be the solution to

$$\rho_z = \exp\left(p_z \binom{n-1}{d-1} (\rho_z^{d-1} - 1)\right). \quad (4.73)$$

Then we know from Theorem 3.1 that $\mathbb{E}[\mathcal{N}(H_d(n, p_z))] = (1 - \rho_z)n$. We do a linear approximation to ρ_z and prove that $n\rho_z \sim n\rho - \lambda_{\mathcal{N}}z$. This is achieved via showing the equality of the first derivative of both sides of (4.73) at $z = 0$ when plugging in $n\rho_z \sim n\rho - \lambda_{\mathcal{N}}z$.

$$\begin{aligned}
 \rho'_z &= \left(\exp(p_z \binom{n-1}{d-1} (\rho_z^{d-1} - 1)) \right)' \\
 \left(\rho - \frac{\lambda_{\mathcal{N}}z}{n} \right)' &= \left(\exp(p_z \binom{n-1}{d-1} ((\rho - \frac{\lambda_{\mathcal{N}}z}{n})^{d-1} - 1)) \right)' \\
 -\frac{\lambda_{\mathcal{N}}}{n} &= \left(p_z \binom{n-1}{d-1} ((\rho - \frac{\lambda_{\mathcal{N}}z}{n})^{d-1} - 1) \right)' \rho_z \\
 &= p'_z \binom{n-1}{d-1} ((\rho - \frac{\lambda_{\mathcal{N}}z}{n})^{d-1} - 1) \rho_z + ((\rho - \frac{\lambda_{\mathcal{N}}z}{n})^{d-1} - 1)' p_z \binom{n-1}{d-1} \rho_z \\
 &= \sigma \binom{n}{d}^{-1} \binom{n-1}{d-1} ((\rho - \frac{\lambda_{\mathcal{N}}z}{n})^{d-1} - 1) \rho_z \\
 &\quad + (-\frac{\lambda_{\mathcal{N}}}{n}(d-1)(\rho - \frac{\lambda_{\mathcal{N}}z}{n})^{d-2}) p_z \binom{n-1}{d-1} \rho_z \\
 &\stackrel{z=0}{=} \sigma \binom{n}{d}^{-1} \binom{n-1}{d-1} (\rho^{d-1} - 1) \rho + (-\frac{\lambda_{\mathcal{N}}}{n}(d-1)\rho^{d-2}) p_0 \binom{n-1}{d-1} \rho \\
 &= \sigma \frac{d}{n} (\rho^{d-1} - 1) \rho + (-\frac{\lambda_{\mathcal{N}}}{n}(d-1)\rho^{d-2}) c \rho \\
 &= -\sigma \frac{d}{n} (1 - \rho^{d-1}) \rho - \frac{\lambda_{\mathcal{N}}}{n} c (d-1) \rho^{d-1} \\
 \lambda_{\mathcal{N}} &= \sigma d (1 - \rho^{d-1}) \rho + \lambda_{\mathcal{N}} c (d-1) \rho^{d-1}
 \end{aligned}$$

which is true as (4.67) shows. Therefore

$$\mathbb{E}[\mathcal{N}(H_d(n, p_z))] \sim (1 - \rho_z)n \sim (1 - \rho)n + \lambda_{\mathcal{N}}z \quad (4.74)$$

and since the small variation with z does not affect the variance:

$$\text{Var}[\mathcal{N}(H_d(n, p_z))] \sim \text{Var}[\mathcal{N}(H_d(n, p_0))] = \sigma_{\mathcal{N}}^2 \quad (4.75)$$

Since we have shown that $\mathcal{N}(H_d(n, p_z))$ has normal distribution (4.74) and (4.75) give

$$\mathbb{P}[\mathcal{N}(H_d(n, p_z)) = (1 - \rho)n + x] = \frac{1}{\sqrt{2\pi}\sigma_{\mathcal{N}}} \exp\left(-\frac{(\lambda_{\mathcal{N}}z - x)^2}{2\sigma_{\mathcal{N}}^2}\right) \quad (4.76)$$

Since the number of edges outside the giant (given the number of vertices outside) is a binomially distributed random variable in $H_d(n, p)$, we can calculate

$$\mathbb{P}[\bar{\mathcal{M}}(H_d(n, p_z)) = \rho^d m_0 - y \mid \mathcal{N}(H_d(n, p_z)) = (1 - \rho)n + x] = \frac{f(z)}{n}. \quad (4.77)$$

First we calculate expectation and variance

$$\begin{aligned}
 \mathbb{E} \left[\bar{\mathcal{M}}(H_d(n, p_z)) \mid \mathcal{N}(H_d(n, p_z)) = (1 - \rho)n + x \right] \\
 &= \binom{\rho n - x}{d} p_z \\
 &\sim \left(\rho^d \binom{n}{d} - x \rho^{d-1} \binom{n-1}{d-1} \right) (p_0 + z \sigma / \binom{n}{d}) \\
 &\sim \rho^d \binom{n}{d} p_0 + \rho^d \sigma z - c \rho^{d-1} x \\
 &\sim \rho^d m_0 + \rho^d \sigma z - c \rho^{d-1} x
 \end{aligned}$$

$$\begin{aligned}
 \text{Var} \left[\bar{\mathcal{M}}(H_d(n, p_z)) \mid \mathcal{N}(H_d(n, p_z)) = (1 - \rho)n + x \right] \\
 &= \binom{\rho n - x}{d} p_z (1 - p_z) \\
 &\sim \left(\rho^d \binom{n}{d} - x \rho^{d-1} \binom{n-1}{d-1} \right) p_0 (1 - p_0) \\
 &\sim \rho^d \sigma^2 - c \rho^{d-1} x \\
 &\sim \rho^d \sigma^2
 \end{aligned}$$

Since the probabilities are independent and the limit of the binomial distribution is a normal one we get for (4.77) (by using (4.76)):

$$f(z) = n \frac{1}{\sqrt{2\pi}\sigma_{\mathcal{N}}} \exp \left(-\frac{(\lambda_{\mathcal{N}} z - x)^2}{2\sigma_{\mathcal{N}}^2} \right) \frac{1}{\sqrt{2\pi\rho^d\sigma^2}} \exp \left(-\frac{(\rho^d \sigma z - c \rho^{d-1} x + y)^2}{2\rho^d \sigma^2} \right)$$

which is the expression used in Lemma 4.5. We will now reformulate this into a constant factor times the density function of a normal distribution in z which will ease the

application of the fourier transform.

$$\begin{aligned}
 \frac{f(z)}{n} &= \frac{1}{\sqrt{2\pi}\sigma_{\mathcal{N}}} \exp\left(-\frac{(\lambda_{\mathcal{N}}z - x)^2}{2\sigma_{\mathcal{N}}^2}\right) \frac{1}{\sqrt{2\pi}\rho^d\sigma^2} \exp\left(-\frac{(\rho^d\sigma z - c\rho^{d-1}x + y)^2}{2\rho^d\sigma^2}\right) \\
 &= \frac{1}{2\pi\sigma_{\mathcal{N}}\rho^{d/2}\sigma} \exp\left(-\frac{1}{2}\left(\left(\frac{\lambda_{\mathcal{N}}}{\sigma_{\mathcal{N}}}\right)^2 z^2 - 2\frac{\lambda_{\mathcal{N}}}{\sigma_{\mathcal{N}}^2}xz + \frac{1}{\sigma_{\mathcal{N}}^2}x^2 \right. \right. \\
 &\quad \left. \left. + \frac{\rho^{2d}\sigma^2 z^2 - 2\rho^d\sigma(c\rho^{d-1}\lambda_{\mathcal{N}}(x-y))z + (c\rho^{d-1}\lambda_{\mathcal{N}}(x-y))^2}{\rho^d\sigma^2}\right)\right) \\
 &= \frac{1}{2\pi\sigma_{\mathcal{N}}\rho^{d/2}\sigma} \exp\left(-\frac{1}{2}\left(\left(\left(\frac{\lambda_{\mathcal{N}}}{\sigma_{\mathcal{N}}}\right)^2 + \rho^d\right)z^2 + 2\left(-\left(\frac{\lambda_{\mathcal{N}}}{\sigma_{\mathcal{N}}}\right)^2 x - \frac{c\rho^{d-1}\lambda_{\mathcal{N}}}{\sigma}(x-y)\right)z \right. \right. \\
 &\quad \left. \left. + \left(\frac{\lambda_{\mathcal{N}}}{\sigma_{\mathcal{N}}}\right)^2 x^2 + \frac{(c\rho^{d-1}\lambda_{\mathcal{N}}(x-y))^2}{\rho^d\sigma^2}\right)\right) \\
 &= \frac{1}{2\pi\sigma_{\mathcal{N}}\rho^{d/2}\sigma} \exp\left(-\frac{1}{2}(\chi z^2 + 2\kappa z + \theta)\right) \\
 &= \frac{1}{2\pi\sigma_{\mathcal{N}}\rho^{d/2}\sigma} \exp\left(-\frac{1}{2}\left(\theta - \frac{\kappa^2}{\chi}\right)\right) \exp\left(-\frac{1}{2}\frac{\left(z + \frac{\kappa}{\chi}\right)^2}{1/\chi}\right) \\
 &= \frac{\exp\left(-\frac{1}{2}\left(\theta - \frac{\kappa^2}{\chi}\right)\right)}{\sqrt{2\pi}\chi\sigma_{\mathcal{N}}\rho^{d/2}\sigma} \frac{1}{\sqrt{2\pi}/\chi} \exp\left(-\frac{1}{2}\frac{\left(z + \frac{\kappa}{\chi}\right)^2}{1/\chi}\right) \\
 &=: \frac{\exp\left(-\frac{1}{2}\left(\theta - \frac{\kappa^2}{\chi}\right)\right)}{\sqrt{2\pi}\chi\sigma_{\mathcal{N}}\rho^{d/2}\sigma} f_1(z)
 \end{aligned}$$

Now we know that f should be the result of the convolution of h with a standard normal distribution and thus get:

$$h_1(z) := \frac{\widehat{f_1}(z)}{\widehat{\phi}(z)} = \frac{1}{\sqrt{2\pi}\sqrt{\frac{1}{\chi} - 1}} \exp\left(-\frac{1}{2}\frac{\left(z + \frac{\kappa}{\chi}\right)^2}{1/\chi - 1}\right) \quad (4.78)$$

which results in

$$\begin{aligned}
 \mathbb{P}\left[\mathcal{N}(H_d(n, m_z)) = (1 - \rho)n + x \wedge \bar{\mathcal{M}}(H_d(n, m_z)) = \rho^d m_0 - y\right] &= \\
 &= \frac{\exp(-\frac{1}{2}(\theta - \frac{\kappa^2}{\chi}))}{\sigma_{\mathcal{N}}\sigma\rho^{d/2}2\pi\sqrt{\chi}\sqrt{\frac{1}{\chi} - 1}} \exp\left(-\frac{\left(z + \frac{\kappa}{\chi}\right)^2}{2(\frac{1}{\chi} - 1)}\right) \quad (4.79)
 \end{aligned}$$

which means by setting $z = 0$

$$\begin{aligned}
 \mathbb{P}\left[\mathcal{N}(H_d(n, m)) = (1 - \rho)n + x \wedge \mathcal{M}(H_d(n, m)) = m(1 - \rho^d) + y\right] &= \\
 &= \frac{1}{\sigma_{\mathcal{N}}\sigma\rho^{d/2}2\pi\sqrt{1 - \chi}} \exp\left(-\frac{1}{2}\left(\theta + \frac{\kappa^2}{1 - \chi}\right)\right) \quad (4.80)
 \end{aligned}$$

We want to bring this in the standard form of a bivariate normal distribution which is

$$\mathbb{P}[\mathcal{N}(H_d(n, m)) = \mu_{\mathcal{N}} + x \wedge \mathcal{M}(H_d(n, m)) = \mu_{\mathcal{M}} + y] = \frac{1}{2\pi\tilde{\sigma}_{\mathcal{N}}\tilde{\sigma}_{\mathcal{M}}\sqrt{1-\tilde{r}^2}} \exp\left(-\frac{1}{2(1-\tilde{r}^2)}\left(\frac{x^2}{\tilde{\sigma}_{\mathcal{N}}^2} - 2\tilde{r}\frac{xy}{\tilde{\sigma}_{\mathcal{N}}\tilde{\sigma}_{\mathcal{M}}} + \frac{y^2}{\tilde{\sigma}_{\mathcal{M}}^2}\right)\right) \quad (4.81)$$

where $\tilde{r} = \frac{\tilde{\sigma}_{\mathcal{N}\mathcal{M}}}{\tilde{\sigma}_{\mathcal{N}}\tilde{\sigma}_{\mathcal{M}}}$, which will result in the values known from Theorem 4.3.

$$\mu_{\mathcal{N}} := (1 - \rho)n \quad (4.82)$$

$$\mu_{\mathcal{M}} := m(1 - \rho^d) \quad (4.83)$$

$$\tilde{\sigma}_{\mathcal{N}}^2 := \sigma_{\mathcal{N}}^2 \left(1 - \left(\frac{\lambda_{\mathcal{N}}}{\sigma_{\mathcal{N}}}\right)^2\right) \quad (4.84)$$

$$\tilde{\sigma}_{\mathcal{M}}^2 := \left(\frac{c\rho^{d-1}\lambda_{\mathcal{N}}\sigma_{\mathcal{N}}}{\lambda_{\mathcal{N}}}\right)^2 - (c\rho^{d-1}\lambda_{\mathcal{N}} - \rho^d\sigma)^2 + \rho^d\sigma^2 \quad (4.85)$$

$$\tilde{r}^2 := \frac{(c\rho^{d-1}\lambda_{\mathcal{N}} - \left(\frac{\lambda_{\mathcal{N}}}{\sigma_{\mathcal{N}}}\right)^2(c\rho^{d-1}\lambda_{\mathcal{N}} - \rho^d\sigma))^2}{\left(1 - \left(\frac{\lambda_{\mathcal{N}}}{\sigma_{\mathcal{N}}}\right)^2\right)\left(\frac{\lambda_{\mathcal{N}}}{\sigma_{\mathcal{N}}}\right)^2\tilde{\sigma}_{\mathcal{M}}^2} \quad (4.86)$$

We show this by proving first

$$\frac{1}{1-\tilde{r}^2} \left(\frac{x^2}{\tilde{\sigma}_{\mathcal{N}}^2} - 2\tilde{r} \frac{xy}{\tilde{\sigma}_{\mathcal{N}}\tilde{\sigma}_{\mathcal{M}}} + \frac{y^2}{\tilde{\sigma}_{\mathcal{M}}^2} \right) = \theta + \frac{\kappa^2}{1-\chi}$$

First we try to separate the terms corresponding to x^2 , xy and y^2 in $\theta + \frac{\kappa^2}{1-\chi}$.

$$\begin{aligned} \theta + \frac{\kappa^2}{1-\chi} &= \left(\frac{\lambda_{\mathcal{N}}}{\sigma_{\mathcal{N}}}\right)^2 x^2 + \frac{(c\rho^{d-1}\lambda_{\mathcal{N}})^2}{\rho^d\sigma^2} (x-y)^2 + \frac{\left(-\left(\frac{\lambda_{\mathcal{N}}}{\sigma_{\mathcal{N}}}\right)^2 x - \frac{c\rho^{d-1}\lambda_{\mathcal{N}}}{\sigma}(x-y)\right)^2}{1 - \left(\frac{\lambda_{\mathcal{N}}}{\sigma_{\mathcal{N}}}\right)^2 - \rho^d} \\ &= \left(\left(\frac{\lambda_{\mathcal{N}}}{\sigma_{\mathcal{N}}}\right)^2 + \frac{(c\rho^{d-1}\lambda_{\mathcal{N}})^2}{\rho^d\sigma^2} + \frac{\left(\left(\frac{\lambda_{\mathcal{N}}}{\sigma_{\mathcal{N}}}\right)^2 + \frac{c\rho^{d-1}\lambda_{\mathcal{N}}}{\sigma}\right)^2}{1 - \left(\frac{\lambda_{\mathcal{N}}}{\sigma_{\mathcal{N}}}\right)^2 - \rho^d} \right) x^2 \\ &\quad - 2 \left(\frac{(c\rho^{d-1}\lambda_{\mathcal{N}})^2}{\rho^d\sigma^2} + \frac{\left(\left(\frac{\lambda_{\mathcal{N}}}{\sigma_{\mathcal{N}}}\right)^2 + \frac{c\rho^{d-1}\lambda_{\mathcal{N}}}{\sigma}\right)\frac{c\rho^{d-1}\lambda_{\mathcal{N}}}{\sigma}}{1 - \left(\frac{\lambda_{\mathcal{N}}}{\sigma_{\mathcal{N}}}\right)^2 - \rho^d} \right) xy \\ &\quad + \left(\frac{(c\rho^{d-1}\lambda_{\mathcal{N}})^2}{\rho^d\sigma^2} + \frac{\left(\frac{c\rho^{d-1}\lambda_{\mathcal{N}}}{\sigma}\right)^2}{1 - \left(\frac{\lambda_{\mathcal{N}}}{\sigma_{\mathcal{N}}}\right)^2 - \rho^d} \right) y^2 \end{aligned}$$

Now it suffices to show the following three equalities (one for each case of x^2 , xy and y^2). We have scaled each equation with a factor of $\lambda_{\mathcal{N}}$ for each x occurring and $c\rho^{d-1}\lambda_{\mathcal{N}}$ for

each y in order to ease calculation.

$$\frac{\lambda_{\mathcal{N}}^2}{(1 - \tilde{r}^2)\tilde{\sigma}_{\mathcal{N}}^2} = \left(\frac{\lambda_{\mathcal{N}}}{\sigma_{\mathcal{N}}}\right)^2 + \frac{(c\rho^{d-1}\lambda_{\mathcal{N}})^2}{\rho^d\sigma^2} + \frac{\left(\left(\frac{\lambda_{\mathcal{N}}}{\sigma_{\mathcal{N}}}\right)^2 + \frac{c\rho^{d-1}\lambda_{\mathcal{N}}}{\sigma}\right)^2}{1 - \left(\frac{\lambda_{\mathcal{N}}}{\sigma_{\mathcal{N}}}\right)^2 - \rho^d} \quad (4.87)$$

$$\frac{\lambda_{\mathcal{N}}c\rho^{d-1}\lambda_{\mathcal{N}}\tilde{r}}{(1 - \tilde{r}^2)\tilde{\sigma}_{\mathcal{N}}\tilde{\sigma}_{\mathcal{M}}} = \frac{(c\rho^{d-1}\lambda_{\mathcal{N}})^2}{\rho^d\sigma^2} + \frac{\left(\left(\frac{\lambda_{\mathcal{N}}}{\sigma_{\mathcal{N}}}\right)^2 + \frac{c\rho^{d-1}\lambda_{\mathcal{N}}}{\sigma}\right)\frac{c\rho^{d-1}\lambda_{\mathcal{N}}}{\sigma}}{1 - \left(\frac{\lambda_{\mathcal{N}}}{\sigma_{\mathcal{N}}}\right)^2 - \rho^d} \quad (4.88)$$

$$\frac{(c\rho^{d-1}\lambda_{\mathcal{N}})^2}{(1 - \tilde{r}^2)\tilde{\sigma}_{\mathcal{M}}^2} = \frac{(c\rho^{d-1}\lambda_{\mathcal{N}})^2}{\rho^d\sigma^2} + \frac{\left(\frac{c\rho^{d-1}\lambda_{\mathcal{N}}}{\sigma}\right)^2}{1 - \left(\frac{\lambda_{\mathcal{N}}}{\sigma_{\mathcal{N}}}\right)^2 - \rho^d} \quad (4.89)$$

A simple auxiliary calculation shows that

$$\begin{aligned} \left(1 - \left(\frac{\lambda_{\mathcal{N}}}{\sigma_{\mathcal{N}}}\right)^2\right) \left(\frac{\lambda_{\mathcal{N}}}{\sigma_{\mathcal{N}}}\right)^2 \tilde{\sigma}_{\mathcal{M}}^2 - (c\rho^{d-1}\lambda_{\mathcal{N}} - \left(\frac{\lambda_{\mathcal{N}}}{\sigma_{\mathcal{N}}}\right)^2 (c\rho^{d-1}\lambda_{\mathcal{N}} - \rho^d\sigma))^2 \\ = \left(\frac{\lambda_{\mathcal{N}}}{\sigma_{\mathcal{N}}}\right)^2 \rho^d\sigma^2 \left(1 - \left(\frac{\lambda_{\mathcal{N}}}{\sigma_{\mathcal{N}}}\right)^2 - \rho^d\right). \end{aligned} \quad (4.90)$$

and now we start with the equality (4.87)

$$\begin{aligned} \frac{\lambda_{\mathcal{N}}^2}{(1 - r^2)\sigma_{\mathcal{N}}^2} &= \frac{\lambda_{\mathcal{N}}^2}{\left(1 - \frac{(c\rho^{d-1}\lambda_{\mathcal{N}} - \left(\frac{\lambda_{\mathcal{N}}}{\sigma_{\mathcal{N}}}\right)^2 (c\rho^{d-1}\lambda_{\mathcal{N}} - \rho^d\sigma))^2}{(1 - \left(\frac{\lambda_{\mathcal{N}}}{\sigma_{\mathcal{N}}}\right)^2) \left(\frac{\lambda_{\mathcal{N}}}{\sigma_{\mathcal{N}}}\right)^2 \sigma_{\mathcal{M}}^2}\right) \sigma_{\mathcal{N}}^2 \left(1 - \left(\frac{\lambda_{\mathcal{N}}}{\sigma_{\mathcal{N}}}\right)^2\right)} \\ &= \frac{\left(\frac{\lambda_{\mathcal{N}}}{\sigma_{\mathcal{N}}}\right)^4 \sigma_{\mathcal{M}}^2}{\left(1 - \left(\frac{\lambda_{\mathcal{N}}}{\sigma_{\mathcal{N}}}\right)^2\right) \left(\frac{\lambda_{\mathcal{N}}}{\sigma_{\mathcal{N}}}\right)^2 \sigma_{\mathcal{M}}^2 - (c\rho^{d-1}\lambda_{\mathcal{N}} - \left(\frac{\lambda_{\mathcal{N}}}{\sigma_{\mathcal{N}}}\right)^2 (c\rho^{d-1}\lambda_{\mathcal{N}} - \rho^d\sigma))^2} \\ &\stackrel{(4.90)}{=} \frac{\left(\frac{\lambda_{\mathcal{N}}}{\sigma_{\mathcal{N}}}\right)^4 \left(\left(\frac{c\rho^{d-1}\lambda_{\mathcal{N}}\sigma_{\mathcal{N}}}{\lambda_{\mathcal{N}}}\right)^2 - (c\rho^{d-1}\lambda_{\mathcal{N}} - \rho^d\sigma)^2 + \rho^d\sigma^2\right)}{\left(\frac{\lambda_{\mathcal{N}}}{\sigma_{\mathcal{N}}}\right)^2 \rho^d\sigma^2 \left(1 - \left(\frac{\lambda_{\mathcal{N}}}{\sigma_{\mathcal{N}}}\right)^2 - \rho^d\right)} \\ &= \frac{(c\rho^{d-1}\lambda_{\mathcal{N}})^2 - \left(\frac{\lambda_{\mathcal{N}}}{\sigma_{\mathcal{N}}}\right)^2 (c\rho^{d-1}\lambda_{\mathcal{N}} - \rho^d\sigma)^2 + \left(\frac{\lambda_{\mathcal{N}}}{\sigma_{\mathcal{N}}}\right)^2 \rho^d\sigma^2}{\rho^d\sigma^2 \left(1 - \left(\frac{\lambda_{\mathcal{N}}}{\sigma_{\mathcal{N}}}\right)^2 - \rho^d\right)} \\ &= \left(\frac{\lambda_{\mathcal{N}}}{\sigma_{\mathcal{N}}}\right)^2 + \frac{(c\rho^{d-1}\lambda_{\mathcal{N}})^2 - \left(\frac{\lambda_{\mathcal{N}}c\rho^{d-1}\lambda_{\mathcal{N}}}{\sigma_{\mathcal{N}}}\right)^2 + 2\left(\frac{\lambda_{\mathcal{N}}}{\sigma_{\mathcal{N}}}\right)^2 c\rho^{d-1}\lambda_{\mathcal{N}}\rho^d\sigma + \left(\frac{\lambda_{\mathcal{N}}}{\sigma_{\mathcal{N}}}\right)^4 \rho^d\sigma^2}{\rho^d\sigma^2 \left(1 - \left(\frac{\lambda_{\mathcal{N}}}{\sigma_{\mathcal{N}}}\right)^2 - \rho^d\right)} \\ &= \left(\frac{\lambda_{\mathcal{N}}}{\sigma_{\mathcal{N}}}\right)^2 + \frac{(c\rho^{d-1}\lambda_{\mathcal{N}})^2}{\rho^d\sigma^2} + \frac{(c\rho^{d-1}\lambda_{\mathcal{N}})^2 \rho^d + 2\left(\frac{\lambda_{\mathcal{N}}}{\sigma_{\mathcal{N}}}\right)^2 c\rho^{d-1}\lambda_{\mathcal{N}}\rho^d\sigma + \left(\frac{\lambda_{\mathcal{N}}}{\sigma_{\mathcal{N}}}\right)^4 \rho^d\sigma^2}{\rho^d\sigma^2 \left(1 - \left(\frac{\lambda_{\mathcal{N}}}{\sigma_{\mathcal{N}}}\right)^2 - \rho^d\right)} \\ &= \left(\frac{\lambda_{\mathcal{N}}}{\sigma_{\mathcal{N}}}\right)^2 + \frac{(c\rho^{d-1}\lambda_{\mathcal{N}})^2}{\rho^d\sigma^2} + \frac{\left(\frac{c\rho^{d-1}\lambda_{\mathcal{N}}}{\sigma} + \left(\frac{\lambda_{\mathcal{N}}}{\sigma_{\mathcal{N}}}\right)^2\right)^2}{1 - \left(\frac{\lambda_{\mathcal{N}}}{\sigma_{\mathcal{N}}}\right)^2 - \rho^d} \end{aligned}$$

and the third (4.89)

$$\begin{aligned}
 \frac{(c\rho^{d-1}\lambda_{\mathcal{N}})^2}{(1-r^2)\sigma_{\mathcal{M}}^2} &= \frac{(c\rho^{d-1}\lambda_{\mathcal{N}})^2}{\left(1 - \frac{(c\rho^{d-1}\lambda_{\mathcal{N}} - (\frac{\lambda_{\mathcal{N}}}{\sigma_{\mathcal{N}}})^2)(c\rho^{d-1}\lambda_{\mathcal{N}} - \rho^d\sigma))^2}{(1 - (\frac{\lambda_{\mathcal{N}}}{\sigma_{\mathcal{N}}})^2)(\frac{\lambda_{\mathcal{N}}}{\sigma_{\mathcal{N}}})^2\sigma_{\mathcal{M}}^2}\right)\sigma_{\mathcal{M}}^2} \\
 &= \frac{(c\rho^{d-1}\lambda_{\mathcal{N}})^2(1 - (\frac{\lambda_{\mathcal{N}}}{\sigma_{\mathcal{N}}})^2)(\frac{\lambda_{\mathcal{N}}}{\sigma_{\mathcal{N}}})^2\sigma_{\mathcal{M}}^2}{((1 - (\frac{\lambda_{\mathcal{N}}}{\sigma_{\mathcal{N}}})^2)(\frac{\lambda_{\mathcal{N}}}{\sigma_{\mathcal{N}}})^2\sigma_{\mathcal{M}}^2 - (c\rho^{d-1}\lambda_{\mathcal{N}} - (\frac{\lambda_{\mathcal{N}}}{\sigma_{\mathcal{N}}})^2(c\rho^{d-1}\lambda_{\mathcal{N}} - \rho^d\sigma))^2)\sigma_{\mathcal{M}}^2} \\
 &\stackrel{(4.90)}{=} \frac{(c\rho^{d-1}\lambda_{\mathcal{N}})^2(1 - (\frac{\lambda_{\mathcal{N}}}{\sigma_{\mathcal{N}}})^2)(\frac{\lambda_{\mathcal{N}}}{\sigma_{\mathcal{N}}})^2}{(\frac{\lambda_{\mathcal{N}}}{\sigma_{\mathcal{N}}})^2\rho^d\sigma^2(1 - (\frac{\lambda_{\mathcal{N}}}{\sigma_{\mathcal{N}}})^2 - \rho^d)} \\
 &= \frac{(c\rho^{d-1}\lambda_{\mathcal{N}})^2(1 - (\frac{\lambda_{\mathcal{N}}}{\sigma_{\mathcal{N}}})^2)}{\rho^d\sigma^2(1 - (\frac{\lambda_{\mathcal{N}}}{\sigma_{\mathcal{N}}})^2 - \rho^d)} \\
 &= \frac{(c\rho^{d-1}\lambda_{\mathcal{N}})^2(1 - (\frac{\lambda_{\mathcal{N}}}{\sigma_{\mathcal{N}}})^2 - \rho^d) + (c\rho^{d-1}\lambda_{\mathcal{N}})^2\rho^d}{\rho^d\sigma^2(1 - (\frac{\lambda_{\mathcal{N}}}{\sigma_{\mathcal{N}}})^2 - \rho^d)} \\
 &= \frac{(c\rho^{d-1}\lambda_{\mathcal{N}})^2}{\rho^d\sigma^2} + \frac{(c\rho^{d-1}\lambda_{\mathcal{N}})^2\rho^d}{\rho^d\sigma^2(1 - (\frac{\lambda_{\mathcal{N}}}{\sigma_{\mathcal{N}}})^2 - \rho^d)} \\
 &= \frac{(c\rho^{d-1}\lambda_{\mathcal{N}})^2}{\rho^d\sigma^2} + \frac{\frac{(c\rho^{d-1}\lambda_{\mathcal{N}})^2}{\sigma^2}}{1 - (\frac{\lambda_{\mathcal{N}}}{\sigma_{\mathcal{N}}})^2 - \rho^d}
 \end{aligned}$$

Now there is still the constant (i.e. independent of x and y) factor in front of the exponential part. It suffices to show that:

$$\tilde{\sigma}_{\mathcal{N}}\tilde{\sigma}_{\mathcal{M}}\sqrt{1-\tilde{r}^2} = \sigma_{\mathcal{N}}\sigma\rho^{d/2}\sqrt{1 - \left(\frac{\lambda_{\mathcal{N}}}{\sigma_{\mathcal{N}}}\right)^2 - \rho^d}$$

We calculate again the squared equation to avoid roots.

$$\begin{aligned}
 \tilde{\sigma}_{\mathcal{N}}^2\tilde{\sigma}_{\mathcal{M}}^2(1-r^2) &= \sigma_{\mathcal{N}}^2(1 - (\frac{\lambda_{\mathcal{N}}}{\sigma_{\mathcal{N}}})^2)\sigma_{\mathcal{M}}^2(1 - \frac{(c\rho^{d-1}\lambda_{\mathcal{N}} - (\frac{\lambda_{\mathcal{N}}}{\sigma_{\mathcal{N}}})^2)(c\rho^{d-1}\lambda_{\mathcal{N}} - \rho^d\sigma))^2}{(1 - (\frac{\lambda_{\mathcal{N}}}{\sigma_{\mathcal{N}}})^2)(\frac{\lambda_{\mathcal{N}}}{\sigma_{\mathcal{N}}})^2\sigma_{\mathcal{M}}^2}) \\
 &= \sigma_{\mathcal{N}}^2(1 - (\frac{\lambda_{\mathcal{N}}}{\sigma_{\mathcal{N}}})^2)\sigma_{\mathcal{M}}^2 - \frac{\sigma_{\mathcal{N}}^2(c\rho^{d-1}\lambda_{\mathcal{N}} - (\frac{\lambda_{\mathcal{N}}}{\sigma_{\mathcal{N}}})^2)(c\rho^{d-1}\lambda_{\mathcal{N}} - \rho^d\sigma))^2}{(\frac{\lambda_{\mathcal{N}}}{\sigma_{\mathcal{N}}})^2} \\
 &= \frac{\sigma_{\mathcal{N}}^4}{\lambda_{\mathcal{N}}^2} \left(\left(\frac{\lambda_{\mathcal{N}}}{\sigma_{\mathcal{N}}}\right)^2 (1 - (\frac{\lambda_{\mathcal{N}}}{\sigma_{\mathcal{N}}})^2)\sigma_{\mathcal{M}}^2 - (c\rho^{d-1}\lambda_{\mathcal{N}} - (\frac{\lambda_{\mathcal{N}}}{\sigma_{\mathcal{N}}})^2(c\rho^{d-1}\lambda_{\mathcal{N}} - \rho^d\sigma))^2 \right) \\
 &\stackrel{(4.90)}{=} \frac{\sigma_{\mathcal{N}}^4}{\lambda_{\mathcal{N}}^2} \left(\frac{\lambda_{\mathcal{N}}}{\sigma_{\mathcal{N}}}\right)^2 \rho^d\sigma^2(1 - (\frac{\lambda_{\mathcal{N}}}{\sigma_{\mathcal{N}}})^2 - \rho^d) \\
 &= \sigma_{\mathcal{N}}^2\rho^d\sigma^2(1 - (\frac{\lambda_{\mathcal{N}}}{\sigma_{\mathcal{N}}})^2 - \rho^d)
 \end{aligned}$$

Variances in terms of c (and ρ)

In order to get the expressions presented in Theorem 4.3 we have to resubstitute (4.63), (4.67), (4.66) into the equations (4.84), (4.85), (4.86):

$$\begin{aligned}
\text{Var} [\mathcal{N}(H_d(n, m))] &= \sigma_{\mathcal{N}}^2 - \lambda_{\mathcal{N}}^2 = \frac{\rho \left(1 - \rho + c(d-1)(\rho - \rho^{d-1})\right) - cd\rho^2(1 - \rho^{d-1})^2}{(1 - c(d-1)\rho^{d-1})^2} n \\
&= \rho \frac{1 - (c+1)\rho - c(d-1)\rho^{d-1} + 2cd\rho^d - cd\rho^{2d-1}}{(1 - c(d-1)\rho^{d-1})^2} n \\
\text{Var} [\mathcal{M}(H_d(n, m))] &= (c\rho^{d-1}\sigma_{\mathcal{N}})^2 - (\lambda_{\mathcal{N}}c\rho^{d-1} - \rho^d\sigma)^2 + \rho^d\sigma^2 \\
&= (c\rho^{d-1}\sigma_{\mathcal{N}})^2 - \frac{cn}{d} \left(\frac{d\rho(1 - \rho^{d-1})}{1 - c(d-1)\rho^{d-1}} c\rho^{d-1} - \rho^d \right)^2 + \frac{cn}{d} \rho^d \\
&= c^2\rho^{2d-2} \frac{\rho \left(1 - \rho + c(d-1)(\rho - \rho^{d-1})\right)}{(1 - c(d-1)\rho^{d-1})^2} n \\
&\quad - \frac{c\rho^{2d}n}{d} \left(\frac{d(1 - \rho^{d-1})}{1 - c(d-1)\rho^{d-1}} c - 1 \right)^2 + \frac{cn}{d} \rho^d \\
\text{Cov} [\mathcal{N}(H_d(n, m)), \mathcal{M}(H_d(n, m))] &= \tilde{r}\tilde{\sigma}_{\mathcal{N}}\tilde{\sigma}_{\mathcal{M}} = \frac{(c\rho^{d-1}\lambda_{\mathcal{N}} - \left(\frac{\lambda_{\mathcal{N}}}{\sigma_{\mathcal{N}}}\right)^2(c\rho^{d-1}\lambda_{\mathcal{N}} - \rho^d\sigma))\sigma_{\mathcal{N}}^2}{\lambda_{\mathcal{N}}} \\
&= c\rho^{d-1}\sigma_{\mathcal{N}}^2 - \lambda_{\mathcal{N}}(\lambda_{\mathcal{N}}c\rho^{d-1} - \rho^d\sigma) \\
&= c\rho^{d-1}(\sigma_{\mathcal{N}}^2 - \lambda_{\mathcal{N}}^2) + \lambda_{\mathcal{N}}\rho^d\sigma \\
&= c\rho^{d-1} \frac{\rho \left(1 - \rho + c(d-1)(\rho - \rho^{d-1})\right) - cd\rho^2(1 - \rho^{d-1})^2}{(1 - c(d-1)\rho^{d-1})^2} n \\
&\quad + \frac{c\rho^{d+1}(1 - \rho^{d-1})}{1 - c(d-1)\rho^{d-1}} n \\
&= c\rho^d \frac{1 - c\rho - c(d-1)\rho^{d-1} + (c + cd - 1)\rho^d - c\rho^{2d-1}}{(1 - c(d-1)\rho^{d-1})^2} n
\end{aligned}$$

For $d = 2$ (the $G_{n,m}$ case) this simplifies to:

$$\begin{aligned}
\text{Var} [\mathcal{N}(G_{n,m})] &= \rho \frac{1 - (2c+1)\rho + 4c\rho^2 - 2c\rho^3}{(1 - c\rho)^2} n \\
&= \rho \frac{1 - \rho - 2c\rho(1 - \rho)^2}{(1 - c\rho)^2} n \\
\text{Var} [\mathcal{M}(G_{n,m})] &= c\rho^2 \frac{1 - (3c^2 - 2c + 1)\rho^2 + 2c(2c - 1)\rho^3 - c^2\rho^4}{2(1 - c\rho)^2} n \\
\text{Cov} [\mathcal{N}(G_{n,m}), \mathcal{M}(G_{n,m})] &= c\rho^2 \frac{1 - 2c\rho + (3c - 1)\rho^2 - c\rho^3}{(1 - c\rho)^2} n
\end{aligned}$$

4.3.2 The Distribution for $H_d(n, p)$

Let

$$\tilde{\chi} := \rho^{-d} - 1 \quad (4.91)$$

$$\tilde{\kappa} := \frac{c\rho^{d-1}\lambda_{\mathcal{N}}}{\rho^d\sigma}(x-y) \quad (4.92)$$

$$\tilde{\theta} := \left(\frac{\lambda_{\mathcal{N}}}{\sigma_{\mathcal{N}}}\right)^2 x^2 + \frac{(c\rho^{d-1}\lambda_{\mathcal{N}})^2}{\rho^d\sigma^2}(x-y)^2 + \frac{\left(\frac{c\rho^{d-1}\lambda_{\mathcal{N}}}{\sigma}y - \left(\frac{\lambda_{\mathcal{N}}}{\sigma_{\mathcal{N}}}\right)^2 + \frac{c\rho^{d-1}\lambda_{\mathcal{N}}}{\sigma}x\right)^2}{1-\chi} \quad (4.93)$$

Now we recalculate the exponential part of (4.79) with respect to $y' = y - z\sigma$ which gives us an explicit approximation for $\tilde{g}(z)$ from (4.9):

$$\begin{aligned} \theta - \frac{\kappa^2}{\chi} + \frac{(z + \frac{\kappa}{\chi})^2}{\frac{1}{\chi} - 1} &= \left(\frac{\lambda_{\mathcal{N}}}{\sigma_{\mathcal{N}}}\right)^2 x^2 + \frac{(c\rho^{d-1}\lambda_{\mathcal{N}}(x-y) + \sigma z)^2}{\rho^d\sigma^2} \\ &\quad - \frac{(-\left(\frac{\lambda_{\mathcal{N}}}{\sigma_{\mathcal{N}}}\right)^2 x - \frac{c\rho^{d-1}\lambda_{\mathcal{N}}}{\sigma}(x-y) - z)^2}{\chi} + \frac{(z + \frac{-\left(\frac{\lambda_{\mathcal{N}}}{\sigma_{\mathcal{N}}}\right)^2 x - \frac{c\rho^{d-1}\lambda_{\mathcal{N}}}{\sigma}(x-y) - z}{\chi})^2}{\frac{1}{\chi} - 1} \\ &= \left(\frac{\lambda_{\mathcal{N}}}{\sigma_{\mathcal{N}}}\right)^2 x^2 + \frac{(c\rho^{d-1}\lambda_{\mathcal{N}})^2(x-y)^2}{\rho^d\sigma^2} + \frac{((\frac{\lambda_{\mathcal{N}}}{\sigma_{\mathcal{N}}})^2 x + \frac{c\rho^{d-1}\lambda_{\mathcal{N}}}{\sigma}(x-y))^2}{1-\chi} \\ &\quad + 2\frac{c\rho^{d-1}\lambda_{\mathcal{N}}(x-y)}{\rho^d\sigma}z + \frac{1}{\rho^d}z^2 + \frac{1-\chi-1}{\chi}z^2 \\ &= \tilde{\chi}z^2 + 2\tilde{\kappa}z + \tilde{\theta} \end{aligned}$$

If we replug this into (4.79) we get:

$$\begin{aligned} \mathbb{P}[\mathcal{N}(H_d(n, m_z)) = (1-\rho)n + x \wedge \bar{\mathcal{M}}(H_d(n, m_z)) = \rho^d m_0 - y] \\ &= \frac{\exp(-\frac{1}{2}(\theta - \frac{\kappa^2}{\chi}))}{\sigma_{\mathcal{N}}\sigma\rho^{d/2}2\pi\sqrt{\chi}\sqrt{\frac{1}{\chi}-1}} \exp\left(-\frac{(z + \frac{\kappa}{\chi})^2}{2(\frac{1}{\chi}-1)}\right) \\ &= \frac{\exp(-\frac{1}{2}(\tilde{\theta} - \frac{\tilde{\kappa}^2}{\tilde{\chi}}))}{\sigma_{\mathcal{N}}\sigma\rho^{d/2}\sqrt{2\pi\chi\tilde{\chi}}\sqrt{\frac{1}{\chi}-1}} \frac{1}{\sqrt{2\pi}\sqrt{\frac{1}{\tilde{\chi}}}} \exp\left(-\frac{(z + \frac{\tilde{\kappa}}{\tilde{\chi}})^2}{2\frac{1}{\tilde{\chi}}}\right) \\ &=: \frac{\exp(-\frac{1}{2}(\tilde{\theta} - \frac{\tilde{\kappa}^2}{\tilde{\chi}}))}{\sigma_{\mathcal{N}}\sigma\rho^{d/2}\sqrt{2\pi\chi\tilde{\chi}}\sqrt{\frac{1}{\chi}-1}} g(z) \end{aligned}$$

This time we need a convolution

$$l(z) := \widehat{\hat{g} \cdot \hat{\phi}}(z) = \frac{1}{\sqrt{2\pi}\sqrt{\frac{1}{\tilde{\chi}}+1}} \exp\left(-\frac{1}{2}\frac{(z + \frac{\tilde{\kappa}}{\tilde{\chi}})^2}{\frac{1}{\tilde{\chi}}+1}\right) \quad (4.94)$$

and get

$$\begin{aligned} \mathbb{P} \left[\mathcal{N}(H_d(n, p_z)) = (1 - \rho)n + x \wedge \mathcal{M}(H_d(n, p_z)) = \frac{cn}{d}(1 - \rho^d) + y \right] \\ = \frac{\exp(-\frac{1}{2}(\tilde{\theta} - \frac{\tilde{\kappa}^2}{\tilde{\chi}}))}{\sigma_{\mathcal{N}}\sigma\rho^{d/2}2\pi\sqrt{\chi\tilde{\chi}}\sqrt{\frac{1}{\chi}-1}\sqrt{\frac{1}{\tilde{\chi}}+1}} \exp \left(-\frac{\left(z + \frac{\tilde{\kappa}}{\tilde{\chi}}\right)^2}{2(\frac{1}{\tilde{\chi}}+1)} \right) \end{aligned}$$

which means

$$\begin{aligned} \mathbb{P} \left[\mathcal{N}(H_d(n, p)) = (1 - \rho)n + x \wedge \mathcal{M}(H_d(n, p)) = \frac{cn}{d}(1 - \rho^d) + y \right] \\ = \frac{1}{\sigma_{\mathcal{N}}\sigma\rho^{d/2}2\pi\sqrt{1-\chi}\sqrt{1+\tilde{\chi}}} \exp \left(-\frac{1}{2} \left(\tilde{\theta} - \frac{\tilde{\kappa}^2}{1+\tilde{\chi}} \right) \right) \end{aligned} \quad (4.95)$$

We want to bring this in the standard form of a bivariate normal distribution (see equation (4.81)) which is indeed possible with

$$\begin{aligned} \mu_{\mathcal{N}} &:= (1 - \rho)n \\ \mu_{\mathcal{M}} &:= \frac{cn}{d}(1 - \rho^d) \\ \sigma_{\mathcal{N}}^2 &:= \sigma_{\mathcal{N}}^2 \end{aligned} \quad (4.96)$$

$$\sigma_{\mathcal{M}}^2 := \left(\frac{c\rho^{d-1}\lambda_{\mathcal{N}}\sigma_{\mathcal{N}}}{\lambda_{\mathcal{N}}} \right)^2 + 2c\rho^{d-1}\lambda_{\mathcal{N}}\sigma + (1 - \rho^d)\sigma^2 \quad (4.97)$$

$$r^2 := \frac{(c\rho^{d-1}\lambda_{\mathcal{N}} + (\frac{\lambda_{\mathcal{N}}}{\sigma_{\mathcal{N}}})^2\sigma)^2}{(\frac{\lambda_{\mathcal{N}}}{\sigma_{\mathcal{N}}})^2\sigma_{\mathcal{M}}^2} \quad (4.98)$$

And once again we want to show that

$$\frac{1}{1 - r^2} \left(\frac{x^2}{\sigma_{\mathcal{N}}^2} - 2r \frac{xy}{\sigma_{\mathcal{N}}\sigma_{\mathcal{M}}} + \frac{y^2}{\sigma_{\mathcal{M}}^2} \right) = \tilde{\theta} - \frac{\tilde{\kappa}^2}{1 + \tilde{\chi}},$$

by separating the three factors to x^2 , xy and y^2 :

$$\begin{aligned}
 \tilde{\theta} - \frac{\tilde{\kappa}^2}{1 + \tilde{\chi}} &= \left(\frac{\lambda_{\mathcal{N}}}{\sigma_{\mathcal{N}}} \right)^2 x^2 + \frac{(c\rho^{d-1}\lambda_{\mathcal{N}})^2}{\rho^d \sigma^2} (x-y)^2 + \frac{\left(\frac{c\rho^{d-1}\lambda_{\mathcal{N}}}{\sigma} y - \left(\left(\frac{\lambda_{\mathcal{N}}}{\sigma_{\mathcal{N}}} \right)^2 + \frac{c\rho^{d-1}\lambda_{\mathcal{N}}}{\sigma} \right) x \right)^2}{1 - \chi} \\
 &\quad - \frac{\left(\frac{c\rho^{d-1}\lambda_{\mathcal{N}}}{\rho^d \sigma} \right)^2 (x-y)^2}{1 + \rho^{-d} - 1} \\
 &= \left(\frac{\lambda_{\mathcal{N}}}{\sigma_{\mathcal{N}}} \right)^2 x^2 + \frac{(c\rho^{d-1}\lambda_{\mathcal{N}})^2}{\rho^d \sigma^2} (x-y)^2 + \frac{\left(\frac{c\rho^{d-1}\lambda_{\mathcal{N}}}{\sigma} y - \left(\left(\frac{\lambda_{\mathcal{N}}}{\sigma_{\mathcal{N}}} \right)^2 + \frac{c\rho^{d-1}\lambda_{\mathcal{N}}}{\sigma} \right) x \right)^2}{1 - \left(\frac{\lambda_{\mathcal{N}}}{\sigma_{\mathcal{N}}} \right)^2 - \rho^d} \\
 &\quad - \frac{(c\rho^{d-1}\lambda_{\mathcal{N}})^2}{\rho^d \sigma^2} (x-y)^2 \\
 &= \left(\frac{\lambda_{\mathcal{N}}}{\sigma_{\mathcal{N}}} \right)^2 x^2 + \frac{\left(\frac{c\rho^{d-1}\lambda_{\mathcal{N}}}{\sigma} y - \left(\left(\frac{\lambda_{\mathcal{N}}}{\sigma_{\mathcal{N}}} \right)^2 + \frac{c\rho^{d-1}\lambda_{\mathcal{N}}}{\sigma} \right) x \right)^2}{1 - \left(\frac{\lambda_{\mathcal{N}}}{\sigma_{\mathcal{N}}} \right)^2 - \rho^d} \\
 &= \left(\left(\frac{\lambda_{\mathcal{N}}}{\sigma_{\mathcal{N}}} \right)^2 + \frac{\left(\left(\frac{\lambda_{\mathcal{N}}}{\sigma_{\mathcal{N}}} \right)^2 + \frac{c\rho^{d-1}\lambda_{\mathcal{N}}}{\sigma} \right)^2}{1 - \left(\frac{\lambda_{\mathcal{N}}}{\sigma_{\mathcal{N}}} \right)^2 - \rho^d} \right) x^2 - 2 \frac{\frac{c\rho^{d-1}\lambda_{\mathcal{N}}}{\sigma} \left(\left(\frac{\lambda_{\mathcal{N}}}{\sigma_{\mathcal{N}}} \right)^2 + \frac{c\rho^{d-1}\lambda_{\mathcal{N}}}{\sigma} \right)}{1 - \left(\frac{\lambda_{\mathcal{N}}}{\sigma_{\mathcal{N}}} \right)^2 - \rho^d} xy \\
 &\quad + \frac{\frac{(c\rho^{d-1}\lambda_{\mathcal{N}})^2}{\sigma^2}}{1 - \left(\frac{\lambda_{\mathcal{N}}}{\sigma_{\mathcal{N}}} \right)^2 - \rho^d} y^2
 \end{aligned}$$

and proving the resulting three equalities

$$\frac{\lambda_{\mathcal{N}}^2}{(1-r^2)\sigma_{\mathcal{N}}^2} = \left(\frac{\lambda_{\mathcal{N}}}{\sigma_{\mathcal{N}}} \right)^2 + \frac{\left(\left(\frac{\lambda_{\mathcal{N}}}{\sigma_{\mathcal{N}}} \right)^2 + \frac{c\rho^{d-1}\lambda_{\mathcal{N}}}{\sigma} \right)^2}{1 - \left(\frac{\lambda_{\mathcal{N}}}{\sigma_{\mathcal{N}}} \right)^2 - \rho^d} \quad (4.99)$$

$$\frac{\lambda_{\mathcal{N}} c \rho^{d-1} \lambda_{\mathcal{N}} r}{(1-r^2)\sigma_{\mathcal{N}}\sigma_{\mathcal{M}}} = \frac{\frac{c\rho^{d-1}\lambda_{\mathcal{N}}}{\sigma} \left(\left(\frac{\lambda_{\mathcal{N}}}{\sigma_{\mathcal{N}}} \right)^2 + \frac{c\rho^{d-1}\lambda_{\mathcal{N}}}{\sigma} \right)}{1 - \left(\frac{\lambda_{\mathcal{N}}}{\sigma_{\mathcal{N}}} \right)^2 - \rho^d} \quad (4.100)$$

$$\frac{(c\rho^{d-1}\lambda_{\mathcal{N}})^2}{(1-r^2)\sigma_{\mathcal{M}}^2} = \frac{\frac{(c\rho^{d-1}\lambda_{\mathcal{N}})^2}{\sigma^2}}{1 - \left(\frac{\lambda_{\mathcal{N}}}{\sigma_{\mathcal{N}}} \right)^2 - \rho^d}. \quad (4.101)$$

Auxiliary calculation shows that

$$\begin{aligned}
 &\left(\frac{\lambda_{\mathcal{N}}}{\sigma_{\mathcal{N}}} \right)^2 \sigma_{\mathcal{M}}^2 - (c\rho^{d-1}\lambda_{\mathcal{N}} + \left(\frac{\lambda_{\mathcal{N}}}{\sigma_{\mathcal{N}}} \right)^2 \sigma)^2 \\
 &= \left(\frac{\lambda_{\mathcal{N}}}{\sigma_{\mathcal{N}}} \right)^2 \sigma^2 (1 - \left(\frac{\lambda_{\mathcal{N}}}{\sigma_{\mathcal{N}}} \right)^2 - \rho^d)
 \end{aligned} \quad (4.102)$$

Equation number one (4.99):

$$\begin{aligned}
\frac{\lambda_{\mathcal{N}}^2}{(1-r^2)\sigma_{\mathcal{N}}^2} &= \frac{\lambda_{\mathcal{N}}^2}{\left(1 - \frac{(c\rho^{d-1}\lambda_{\mathcal{N}} + (\frac{\lambda_{\mathcal{N}}}{\sigma_{\mathcal{N}}})^2\sigma)^2}{(\frac{\lambda_{\mathcal{N}}}{\sigma_{\mathcal{N}}})^2\sigma_{\mathcal{M}}^2}\right)\sigma_{\mathcal{N}}^2} \\
&= \frac{(\frac{\lambda_{\mathcal{N}}}{\sigma_{\mathcal{N}}})^4\sigma_{\mathcal{M}}^2}{(\frac{\lambda_{\mathcal{N}}}{\sigma_{\mathcal{N}}})^2\sigma_{\mathcal{M}}^2 - (c\rho^{d-1}\lambda_{\mathcal{N}} + (\frac{\lambda_{\mathcal{N}}}{\sigma_{\mathcal{N}}})^2\sigma)^2} \\
&\stackrel{(4.102)}{=} \frac{(\frac{\lambda_{\mathcal{N}}}{\sigma_{\mathcal{N}}})^4 \left(\left(\frac{c\rho^{d-1}\lambda_{\mathcal{N}}\sigma_{\mathcal{N}}}{\lambda_{\mathcal{N}}} \right)^2 + 2c\rho^{d-1}\lambda_{\mathcal{N}}\sigma + (1-\rho^d)\sigma^2 \right)}{(\frac{\lambda_{\mathcal{N}}}{\sigma_{\mathcal{N}}})^2\sigma^2(1 - (\frac{\lambda_{\mathcal{N}}}{\sigma_{\mathcal{N}}})^2 - \rho^d)} \\
&= \frac{\left(\frac{c\rho^{d-1}\lambda_{\mathcal{N}}}{\sigma} + (\frac{\lambda_{\mathcal{N}}}{\sigma_{\mathcal{N}}})^2 \right)^2 + (\frac{\lambda_{\mathcal{N}}}{\sigma_{\mathcal{N}}})^2(1 - (\frac{\lambda_{\mathcal{N}}}{\sigma_{\mathcal{N}}})^2 - \rho^d)}{1 - (\frac{\lambda_{\mathcal{N}}}{\sigma_{\mathcal{N}}})^2 - \rho^d} \\
&= \left(\frac{\lambda_{\mathcal{N}}}{\sigma_{\mathcal{N}}} \right)^2 + \frac{\left(\frac{c\rho^{d-1}\lambda_{\mathcal{N}}}{\sigma} + (\frac{\lambda_{\mathcal{N}}}{\sigma_{\mathcal{N}}})^2 \right)^2}{1 - (\frac{\lambda_{\mathcal{N}}}{\sigma_{\mathcal{N}}})^2 - \rho^d}
\end{aligned}$$

Equation number two (4.100):

$$\begin{aligned}
\frac{\lambda_{\mathcal{N}}c\rho^{d-1}\lambda_{\mathcal{N}}r}{(1-r^2)\sigma_{\mathcal{N}}\sigma_{\mathcal{M}}} &= \frac{\lambda_{\mathcal{N}}c\rho^{d-1}\lambda_{\mathcal{N}}\frac{c\rho^{d-1}\lambda_{\mathcal{N}} + (\frac{\lambda_{\mathcal{N}}}{\sigma_{\mathcal{N}}})^2\sigma}{\frac{\lambda_{\mathcal{N}}}{\sigma_{\mathcal{N}}}\sigma_{\mathcal{M}}}}{\left(1 - \frac{(c\rho^{d-1}\lambda_{\mathcal{N}} + (\frac{\lambda_{\mathcal{N}}}{\sigma_{\mathcal{N}}})^2\sigma)^2}{(\frac{\lambda_{\mathcal{N}}}{\sigma_{\mathcal{N}}})^2\sigma_{\mathcal{M}}^2}\right)\sigma_{\mathcal{N}}\sigma_{\mathcal{M}}} \\
&= \frac{(\frac{\lambda_{\mathcal{N}}}{\sigma_{\mathcal{N}}})^2\sigma_{\mathcal{M}}c\rho^{d-1}\lambda_{\mathcal{N}}(c\rho^{d-1}\lambda_{\mathcal{N}} + (\frac{\lambda_{\mathcal{N}}}{\sigma_{\mathcal{N}}})^2\sigma)}{((\frac{\lambda_{\mathcal{N}}}{\sigma_{\mathcal{N}}})^2\sigma_{\mathcal{M}}^2 - (c\rho^{d-1}\lambda_{\mathcal{N}} + (\frac{\lambda_{\mathcal{N}}}{\sigma_{\mathcal{N}}})^2\sigma)^2)\sigma_{\mathcal{M}}} \\
&= \frac{(\frac{\lambda_{\mathcal{N}}}{\sigma_{\mathcal{N}}})^2c\rho^{d-1}\lambda_{\mathcal{N}}(c\rho^{d-1}\lambda_{\mathcal{N}} + (\frac{\lambda_{\mathcal{N}}}{\sigma_{\mathcal{N}}})^2\sigma)}{(\frac{\lambda_{\mathcal{N}}}{\sigma_{\mathcal{N}}})^2\sigma_{\mathcal{M}}^2 - (c\rho^{d-1}\lambda_{\mathcal{N}} + (\frac{\lambda_{\mathcal{N}}}{\sigma_{\mathcal{N}}})^2\sigma)^2} \\
&\stackrel{(4.102)}{=} \frac{(\frac{\lambda_{\mathcal{N}}}{\sigma_{\mathcal{N}}})^2c\rho^{d-1}\lambda_{\mathcal{N}}(c\rho^{d-1}\lambda_{\mathcal{N}} + (\frac{\lambda_{\mathcal{N}}}{\sigma_{\mathcal{N}}})^2\sigma)}{(\frac{\lambda_{\mathcal{N}}}{\sigma_{\mathcal{N}}})^2\sigma^2(1 - (\frac{\lambda_{\mathcal{N}}}{\sigma_{\mathcal{N}}})^2 - \rho^d)} \\
&= \frac{c\rho^{d-1}\lambda_{\mathcal{N}}(c\rho^{d-1}\lambda_{\mathcal{N}} + (\frac{\lambda_{\mathcal{N}}}{\sigma_{\mathcal{N}}})^2\sigma)}{\sigma^2(1 - (\frac{\lambda_{\mathcal{N}}}{\sigma_{\mathcal{N}}})^2 - \rho^d)} \\
&= \frac{\frac{c\rho^{d-1}\lambda_{\mathcal{N}}}{\sigma} \left(\frac{c\rho^{d-1}\lambda_{\mathcal{N}}}{\sigma} + (\frac{\lambda_{\mathcal{N}}}{\sigma_{\mathcal{N}}})^2 \right)}{1 - (\frac{\lambda_{\mathcal{N}}}{\sigma_{\mathcal{N}}})^2 - \rho^d}
\end{aligned}$$

Equation number three (4.101):

$$\begin{aligned}
 \frac{(c\rho^{d-1}\lambda_{\mathcal{N}})^2}{(1-r^2)\sigma_{\mathcal{M}}^2} &= \frac{(c\rho^{d-1}\lambda_{\mathcal{N}})^2}{\left(1 - \frac{(c\rho^{d-1}\lambda_{\mathcal{N}} + (\frac{\lambda_{\mathcal{N}}}{\sigma_{\mathcal{N}}})^2\sigma)^2}{(\frac{\lambda_{\mathcal{N}}}{\sigma_{\mathcal{N}}})^2\sigma_{\mathcal{M}}^2}\right)\sigma_{\mathcal{M}}^2} \\
 &= \frac{(c\rho^{d-1}\lambda_{\mathcal{N}})^2(\frac{\lambda_{\mathcal{N}}}{\sigma_{\mathcal{N}}})^2\sigma_{\mathcal{M}}^2}{((\frac{\lambda_{\mathcal{N}}}{\sigma_{\mathcal{N}}})^2\sigma_{\mathcal{M}}^2 - (c\rho^{d-1}\lambda_{\mathcal{N}} + (\frac{\lambda_{\mathcal{N}}}{\sigma_{\mathcal{N}}})^2\sigma)^2)\sigma_{\mathcal{M}}^2} \\
 &= \frac{(c\rho^{d-1}\lambda_{\mathcal{N}})^2(\frac{\lambda_{\mathcal{N}}}{\sigma_{\mathcal{N}}})^2}{(\frac{\lambda_{\mathcal{N}}}{\sigma_{\mathcal{N}}})^2\sigma_{\mathcal{M}}^2 - (c\rho^{d-1}\lambda_{\mathcal{N}} + (\frac{\lambda_{\mathcal{N}}}{\sigma_{\mathcal{N}}})^2\sigma)^2} \\
 &\stackrel{(4.102)}{=} \frac{(c\rho^{d-1}\lambda_{\mathcal{N}})^2(\frac{\lambda_{\mathcal{N}}}{\sigma_{\mathcal{N}}})^2}{(\frac{\lambda_{\mathcal{N}}}{\sigma_{\mathcal{N}}})^2\sigma^2(1 - (\frac{\lambda_{\mathcal{N}}}{\sigma_{\mathcal{N}}})^2 - \rho^d)} \\
 &= \frac{\frac{(c\rho^{d-1}\lambda_{\mathcal{N}})^2}{\sigma^2}}{1 - (\frac{\lambda_{\mathcal{N}}}{\sigma_{\mathcal{N}}})^2 - \rho^d}
 \end{aligned}$$

Last but not least we need to show

$$\begin{aligned}
 \sigma_{\mathcal{N}}\sigma_{\mathcal{M}}\sqrt{1-r^2} &= \sigma_{\mathcal{N}}\sigma\rho^{d/2}\sqrt{1-\chi}\sqrt{1+\tilde{\chi}} \\
 &= \sigma_{\mathcal{N}}\sigma\rho^{d/2}\sqrt{1 - \left(\frac{\lambda_{\mathcal{N}}}{\sigma_{\mathcal{N}}}\right)^2 - \rho^d}\sqrt{1 + \rho^{-d} - 1} \\
 &= \sigma_{\mathcal{N}}\sigma\sqrt{1 - \left(\frac{\lambda_{\mathcal{N}}}{\sigma_{\mathcal{N}}}\right)^2 - \rho^d}
 \end{aligned}$$

We calculate again the squared equation to avoid roots.

$$\begin{aligned}
 \sigma_{\mathcal{N}}^2\sigma_{\mathcal{M}}^2(1-r^2) &= \sigma_{\mathcal{N}}^2\sigma_{\mathcal{M}}^2\left(1 - \frac{(c\rho^{d-1}\lambda_{\mathcal{N}} + (\frac{\lambda_{\mathcal{N}}}{\sigma_{\mathcal{N}}})^2\sigma)^2}{(\frac{\lambda_{\mathcal{N}}}{\sigma_{\mathcal{N}}})^2\sigma_{\mathcal{M}}^2}\right) \\
 &= \frac{\sigma_{\mathcal{N}}^4}{\lambda_{\mathcal{N}}^2}\left(\left(\frac{\lambda_{\mathcal{N}}}{\sigma_{\mathcal{N}}}\right)^2\sigma_{\mathcal{M}}^2 - (c\rho^{d-1}\lambda_{\mathcal{N}} + \left(\frac{\lambda_{\mathcal{N}}}{\sigma_{\mathcal{N}}}\right)^2\sigma)^2\right) \\
 &\stackrel{(4.102)}{=} \frac{\sigma_{\mathcal{N}}^4}{\lambda_{\mathcal{N}}^2}\left(\frac{\lambda_{\mathcal{N}}}{\sigma_{\mathcal{N}}}\right)^2\sigma^2\left(1 - \left(\frac{\lambda_{\mathcal{N}}}{\sigma_{\mathcal{N}}}\right)^2 - \rho^d\right) \\
 &= \sigma_{\mathcal{N}}^2\sigma^2\left(1 - \left(\frac{\lambda_{\mathcal{N}}}{\sigma_{\mathcal{N}}}\right)^2 - \rho^d\right) \tag{4.103}
 \end{aligned}$$

Variances in terms of c (and ρ)

Resubstituting (4.63), (4.67), (4.66) into the equations (4.96), (4.97), (4.98) gives:

$$\begin{aligned}
\text{Var} [\mathcal{N}(H_d(n, p))] &= \frac{\rho \left(1 - \rho + c(d-1)(\rho - \rho^{d-1})\right)}{(1 - c(d-1)\rho^{d-1})^2} n \\
\text{Var} [\mathcal{M}(H_d(n, p))] &= (c\rho^{d-1}\sigma_{\mathcal{N}})^2 + 2\lambda_{\mathcal{N}}c\rho^{d-1}\sigma + (1 - \rho^d)\sigma^2 \\
&= c^2\rho^{2d-2} \frac{\rho \left(1 - \rho + c(d-1)(\rho - \rho^{d-1})\right)}{(1 - c(d-1)\rho^{d-1})^2} n + 2 \frac{c^2\rho^d(1 - \rho^{d-1})}{1 - c(d-1)\rho^{d-1}} n \\
&\quad + (1 - \rho^d) \frac{cn}{d} \\
&= c^2\rho^d \frac{2 + c(d-1)\rho^{2d-2} - 2c(d-1)\rho^{d-1} + c(d-1)\rho^d - \rho^{d-1} - \rho^d}{(1 - c(d-1)\rho^{d-1})^2} n \\
&\quad + (1 - \rho^d) \frac{cn}{d}
\end{aligned}$$

$$\begin{aligned}
\text{Cov} [\mathcal{N}(H_d(n, p)), \mathcal{M}(H_d(n, p))] &= r\sigma_{\mathcal{N}}\sigma_{\mathcal{M}} = \frac{(c\rho^{d-1}\lambda_{\mathcal{N}} + (\frac{\lambda_{\mathcal{N}}}{\sigma_{\mathcal{N}}})^2\sigma)\sigma_{\mathcal{N}}^2}{\lambda_{\mathcal{N}}} \\
&= c\rho^{d-1}\sigma_{\mathcal{N}}^2 + \lambda_{\mathcal{N}}\sigma \\
&= c\rho^{d-1} \frac{\rho \left(1 - \rho + c(d-1)(\rho - \rho^{d-1})\right)}{(1 - c(d-1)\rho^{d-1})^2} n \\
&\quad + \frac{c\rho(1 - \rho^{d-1})}{1 - c(d-1)\rho^{d-1}} n \\
&= c\rho \frac{1 - \rho^d - c(d-1)\rho^{d-1}(1 - \rho)}{(1 - c(d-1)\rho^{d-1})^2} n
\end{aligned}$$

and for $d = 2$:

$$\begin{aligned}
\text{Var} [\mathcal{N}(G_{n,p})] &= \frac{\rho(1 - \rho)}{(1 - c\rho)^2} n \\
\text{Var} [\mathcal{M}(G_{n,p})] &= c^2\rho^2 \frac{2 + (2c-1)\rho^2 - (2c+1)\rho}{(1 - c\rho)^2} n + (1 - \rho^2) \frac{cn}{2} \\
\text{Cov} [\mathcal{N}(G_{n,p}), \mathcal{M}(G_{n,p})] &= c\rho \frac{1 - \rho^2 - c\rho(1 - \rho)}{(1 - c\rho)^2} n
\end{aligned}$$

Chapter 5

Applications of the Local Limit Theorems

5.1 Results

5.1.1 The Probability of Connectedness

As an application of the local limit theorem for $H_d(n, p)$ (Theorem 4.1), we obtain the following formula for the asymptotic probability that $H_d(n, m)$ is connected, and thus for the number of connected hypergraphs of a given order and size.

Theorem 5.1. *Let $d \geq 2$ be a fixed integer. For any compact set $\mathcal{J} \subset (d/(d-1), \infty)$ the following holds. Let $m = m(n)$ be a sequence of integers such that $c = c(n) = dm/n \in \mathcal{J}$ for all n and let $0 < \rho = \rho(n) < 1$ be the unique solution to*

$$\rho_m = \exp \left(-c(1 - \rho_m) \cdot \frac{1 - \rho_m^{d-1}}{1 - \rho_m^d} \right). \quad (5.1)$$

Furthermore let

$$\Phi_d(c) = \rho_m^{\frac{\rho_m}{1-\rho_m}} (1 - \rho_m) \left(\frac{1 - \rho_m^d}{(1 - \rho_m)^d} \right)^{\frac{c}{d}}.$$

and let $c_d(n, m)$ denote the probability that $H_d(n, m)$ is connected while $C_d(n, m)$ denotes the number of connected d -uniform hypergraphs of order n and size m . If $n \geq n_0$ then for $d = 2$:

$$\begin{aligned} c_2(n, m) &= C_2(n, m) \left(\frac{\binom{n}{2}}{m} \right)^{-1} \\ &\sim \frac{1 + \rho_m - c\rho_m}{\sqrt{(1 + \rho_m)^2 - 2c\rho_m}} \exp \left(\frac{2c\rho_m + c^2\rho_m}{2(1 + \rho_m)} \right) \cdot \Phi_2(\rho_m, c)^n \end{aligned}$$

and for $d > 2$:

$$\begin{aligned}
 c_d(n, m) &= C_d(n, m) \binom{n}{d}^{-1} \\
 &\sim \frac{1 - \rho_m^d - (1 - \rho_m)c(d-1)\rho_m^{d-1}}{\sqrt{(1 - \rho_m^d + c(d-1)(\rho_m - \rho_m^{d-1}))(1 - \rho_m^d) - cd\rho_m(1 - \rho_m^{d-1})^2}} \\
 &\quad \cdot \exp\left(\frac{c(d-1)(\rho_m - 2\rho_m^d + \rho_m^{d-1})}{2(1 - \rho_m^d)}\right) \cdot \Phi_d(\rho_m, c)^n
 \end{aligned}$$

To prove Theorem 5.1 we actually need the *local* limit theorem for $\mathcal{N}, \mathcal{M}(H_d(n, p))$; that is, Theorem 5.1 cannot be derived from just the central limit theorem provided by Corollary 4.2.

Furthermore, we have the following theorem on the asymptotic probability that $H_d(n, p)$ is connected.

Theorem 5.2. *Let $d \geq 2$ be a fixed integer. For any compact set $\mathcal{J} \subset (0, \infty)$ the following holds. Let $p = p(n)$ be a sequence such that $c = c(n) = \binom{n-1}{d-1}p \in \mathcal{J}$ for all n and let $0 < \varrho = \varrho(n) < 1$ be the unique solution to*

$$\varrho = \exp\left(c \frac{\varrho^{d-1} - 1}{(1 - \varrho)^{d-1}}\right). \quad (5.2)$$

If we let $c_d(n, p) = \mathbb{P}[H_d(n, p) \text{ is connected}]$ and

$$\Psi_d(c) = \varrho^{\frac{c}{1-\varrho}} (1 - \varrho) \exp\left(\frac{c}{d} \frac{1 - \varrho^d - (1 - \varrho)^d}{(1 - \varrho)^d}\right)$$

and if $n \geq n_0$ then for $d = 2$:

$$\begin{aligned}
 c_2(n, p) &\sim \exp\left(\frac{2ce^{-c} + 2c + c^2}{2(e^c - 1)}\right) \left(1 - \frac{c}{e^c - 1}\right) \Psi_2(c)^n \\
 &= \exp\left(\frac{2ce^{-c} + 2c + c^2}{2(e^c - 1)}\right) \left(1 - \frac{c}{e^c - 1}\right) (1 - e^{-c})^n
 \end{aligned}$$

and for $d > 2$:

$$\begin{aligned}
 c_d(n, p) &\sim \exp\left(c\varrho \frac{1 - \varrho^d - (1 - \varrho)^d}{d(1 - \varrho)^d} + \frac{c(d-1)\varrho}{2} \left(\left(\frac{\varrho}{1 - \varrho}\right)^{d-2} + 1\right)\right) \\
 &\quad \cdot \frac{1 - c(d-1)\left(\frac{\varrho}{1 - \varrho}\right)^{d-1}}{\sqrt{1 + c(d-1)\frac{\varrho - \varrho^{d-1}}{(1 - \varrho)^d}}} \cdot \Psi_d(c)^n.
 \end{aligned}$$

Interestingly, if we choose $p = p(n)$ and $m = m(n)$ in such a way that $\binom{n}{d}p = m$ for all n and set $c(n) = \binom{n-1}{d-1}p = dm/n$, then the function $\Psi_d(c)$ is strictly bigger than $\Phi_d(c)$ for all values of c . Consequently, the probability that $H_d(n, p)$ is connected exceeds the probability that $H_d(n, m)$ is connected by an exponential factor. The reason is that in $H_d(n, p)$ the total number of edges is a (binomially distributed) random variable. In fact, it turns out that – roughly speaking – $H_d(n, p)$ “boosts” its probability of connectivity by including a number of edges that exceeds $\binom{n}{d}p$ considerably. That is, the total number of edges *given that $H_d(n, p)$ is connected* is significantly bigger than $\binom{n}{d}p$.

5.1.2 The Distribution of $\mathcal{M}(H_d(n, p))$ given Connectedness

The following local limit theorem for the total number of edges in $H_d(n, p)$ given that $H_d(n, p)$ is connected quantifies this observation.

Theorem 5.3. *Let $d \geq 2$ be a fixed integer. For any two compact sets $\mathcal{I} \subset \mathbb{R}$, $\mathcal{J} \subset (0, \infty)$ the following holds. Let $p = p(n)$ be a sequence such that $c = c(n) = \binom{n-1}{d-1}p \in \mathcal{J}$ for all n and let $0 < \varrho = \varrho(n) < 1$ be the unique solution to (5.2). If $n \geq n_0$ and if y is such that $n^{-\frac{1}{2}}y \in \mathcal{I}$, then*

$$\mathbb{P}[|E(H_d(n, p))| = \mu_g + y \mid H_d(n, p) \text{ is connected}] \sim \frac{1}{\sqrt{2\pi}\sigma_g} \exp\left(-\frac{y^2}{2\sigma_g^2}\right)$$

with $\mu_g = \frac{cn}{d(1-\varrho)^d}(1 - \varrho^d)$ and $\sigma_g^2 = \frac{cn}{d(1-\varrho)^d}\left(1 - \frac{cd\varrho(1-\varrho^{d-1})^2}{(1-\varrho)^d + c(d-1)(\varrho - \varrho^{d-1})} - \varrho^d\right)$.

5.2 Techniques

First we state two corollaries on the probability that order and size of the giant component are very close to their expected values, which follow directly from Theorems 4.1 and 4.3 of the last chapter.

Corollary 5.4. *For each $c > c_0 > (d-1)^{-1}$ with c_0 being constant and $p = c/\binom{n-1}{d-1}$ let $\rho = \rho(c)$ be the single solution of $\rho = e^{c(\rho^{d-1}-1)}$ in the interval $[0, 1]$. Then for every constant k the following holds*

$$\begin{aligned} \mathbb{P}[\mathcal{N}(H_d(n, p)) = (1 - \rho)n + k] &\sim (2\pi \text{Var}[\mathcal{N}(H_d(n, p))])^{-1/2} \\ &\sim \left(2\pi n \frac{\rho(1 - \rho + c(d-1)(\rho - \rho^{d-1}))}{(1 - c(d-1)\rho^{d-1})^2}\right)^{-1/2} \end{aligned}$$

For $d = 2$ this simplifies to

$$\mathbb{P}[\mathcal{N}(H_d(n, p)) = (1 - \rho)n + k] \sim \left(2\pi n \frac{\rho(1 - \rho)}{(1 - c\rho)^2}\right)^{-1/2}$$

Corollary 5.5. *For each $c > c_0 > (d-1)^{-1}$ with c_0 being constant and $p = c/\binom{n-1}{d-1}$ let $\rho = \rho(c)$ be the single solution of $\rho = e^{c(\rho^{d-1}-1)}$ in the interval $[0, 1)$. Then for every constant k the following holds*

$$\mathbb{P} \left[\mathcal{N}(H_d(n, p)) = (1 - \rho)n + k \wedge \mathcal{M}(H_d(n, p)) = (1 - \rho^d) \frac{cn}{d} + y \right] \sim$$

$$\frac{1}{2\pi n} \frac{1 - c(d-1)\rho^{d-1}}{\sqrt{\frac{c}{d}\rho(1 - \rho + c(d-1)(\rho - \rho^{d-1}))((1 - \rho^d) - c^2\rho^2(1 - \rho^{d-1})^2)}} \cdot \exp \left(-\frac{y^2 d}{2cn} \left(1 - \frac{cd\rho(1 - \rho^{d-1})^2}{1 - \rho + c(d-1)(\rho - \rho^{d-1})} - \rho^d \right)^{-1} \right)$$

All of the proofs use the Corollaries 5.4 and 5.5 together with the fact that the giant component is a uniform random connected graph (supposed the number of vertices is given). In order to use this result we will often need to have the giant component of a graph to have a certain number of vertices and edges and thus fix parameters n and p for a random hypergraph $H_d(n, p)$ to fulfil this assumptions.

One problem occurring multiple times in this context are integrality issues which will need a lot of effort in the forthcoming sections. One single lemma which will be helpful in solving these problems shall be stated here:

Lemma 5.6. *Let $c_1, c_2 > 1/(d-1)$ with $|c_1 - c_2| = O(\frac{1}{n})$ and let ρ_i be the solution to $\rho_i = 1 - \exp(c_i(\rho_i^{d-1} - 1))$ for $i = 1, 2$. Then $|\rho_1 - \rho_2| = O(\frac{1}{n})$*

Proof. Calculate the derivative and use the Taylor series expansion, which is possible because ρ is differentiable due to the implicit function theorem. \square

Since a random graph $H_d(n, p)$ with $p = c/\binom{n-1}{d-1}$ has essentially a giant component of order $(1 - \rho)n$ with ρ being as in Lemma 5.6 this lemma allows for small variations in c (or p and n respectively) without affecting the value of ρ too much. Although we did state the theorems in terms of n , m , and p we will in the proofs reserve these letters for talking about the graph whose giant component we analyse (where the giant component is the graph the theorems are about). If we talk about order and size of the giant (and sometimes a slightly altered probability to solve integrality issues) we use the letters ν , μ , and q . Furthermore we replace $c = p\binom{n-1}{d-1}$ by $\zeta = q\binom{\nu-1}{d-1}$.

5.3 Probability of Connectedness in the Binomial Model

The proof will follow mainly the lines of the proof of Theorem 2 in Coja-Oghlan et al. [2006] while using additional ideas from the proof of Theorem 1 in Coja-Oghlan et al. [2006].

For given numbers ν and ζ with $\zeta > c_0$ for a constant $c_0 > 0$ we want to compute the probability that $H_d(\nu, p)$ is connected where $p = \zeta/\binom{\nu-1}{d-1}$. We choose n such that the giant component of $H_d(n, p)$ is close to $H_d(\nu, p)$.

For any integer n such that $c = p \binom{n-1}{d-1} > (d-1)^{-1}$, the equation $\rho = \exp(c(\rho^{d-1} - 1))$ has a unique solution $0 < \rho < 1$. Now, let n be the largest integer such that $(1 - \rho)n \leq \nu$, furthermore let ρ' be such that $(1 - \rho')n = \nu$. These definitions of ν , ζ , n , c , ρ , and ρ' are valid in the whole section while ϱ is defined as in (5.2).

Lemma 5.7. $\nu - (1 - \rho)n = O(1)$

Proof. It is clear that the function $\rho(c)$ defined by $\rho = \exp(c(\rho^{d-1} - 1))$ is continuous in c thus the only problem is that c cannot take any value because it is defined by $c = p \binom{n-1}{d-1}$ for discrete n and d and a given $p = \zeta / \binom{\nu-1}{d-1}$. In the light of Lemma 5.6 it suffices to show that $c(n+1) - c(n) = O(\frac{1}{n})$.

$$\begin{aligned} p \binom{n}{d-1} - p \binom{n-1}{d-1} &= p \binom{n-1}{d-1} \left(\frac{n}{n-d+1} - 1 \right) \\ &= \frac{\zeta \binom{n-1}{d-1} \frac{d-1}{n-d+1}}{\binom{\nu-1}{d-1}} \\ &= O\left(\frac{1}{n}\right) \end{aligned}$$

□

Lemma 5.8.

$$c = \zeta(1 - \rho')^{1-d} \left(1 + \binom{d}{2} \frac{\rho'}{(1 - \rho')n} + O(n^{-2}) \right)$$

Proof. Since $c = \zeta \binom{n-1}{d-1} / \binom{\nu-1}{d-1}$ it suffices to show that

$$\begin{aligned} (1 - \rho')^{d-1} \frac{\binom{n-1}{d-1}}{\binom{(1-\rho')^{n-1}}{d-1}} &= \frac{(1 - \rho')^{d-1} (n-1)_{d-1}}{((1 - \rho')n - 1)_{d-1}} = \prod_{j=1}^{d-1} \frac{(1 - \rho')(n-j)}{(1 - \rho')n - j} \\ &= \prod_{j=1}^{d-1} 1 + \frac{\rho'j}{(1 - \rho')n - j} \\ &= \exp \left(\sum_{j=1}^{d-1} \frac{\rho'j}{(1 - \rho')n} + O(n^{-2}) \right) \\ &= \exp \left(\binom{d}{2} \frac{\rho'}{(1 - \rho')n} + O(n^{-2}) \right). \end{aligned}$$

□

Lemma 5.9.

$$\mathbb{P}[\mathcal{N}(H_d(n, p)) = \nu] \sim \binom{n}{\nu} c_d(\nu, p) (1-p)^{\binom{n}{d} - \binom{n-\nu}{d} - \binom{\nu}{d}}$$

Proof. The right hand side denotes just the expected number of components of order ν occurring in $H_d(n, p)$, and thus provides an upper bound on $\mathbb{P}[\mathcal{N}(H_d(n, p)) = \nu]$. The three factor denote the number of ways to choose a set of ν vertices where to place a component of size ν , the probability that this subhypergraph of $H_d(n, p)$ is connected ($c_d(\nu, p)$) and the probability that none of the $\binom{n}{d} - \binom{\nu}{d} - \binom{n-\nu}{d}$ possible edges connecting the set of size ν with the rest of the graph is present in $H_d(n, p)$, and each of these edges occurs with probability p independently. On the other hand, we have

$$\mathbb{P}[\mathcal{N}(H_d(n, p)) = \nu] \geq \binom{n}{\nu} c_d(\nu, p) (1-p)^{\binom{n}{d} - \binom{\nu}{d} - \binom{n-\nu}{d}} \mathbb{P}[\mathcal{N}(H_d(n-\nu, p)) < \nu],$$

because the term on the right hand side equals the probability that there is precisely one component of order ν . As $\mathbb{P}[\mathcal{N}(H_d(n-\nu, p)) < \nu] \sim 1$ by [Coja-Oghlan et al., 2006, Lemma 9], the statement follows. \square

Lemma 5.10. *Let*

$$\Psi_d(x, \zeta) = (1-x)x^{\frac{x}{1-x}} \exp\left(\frac{\zeta}{d} \frac{1-x^d - (1-x)^d}{(1-x)^d}\right)$$

Then $\Psi_d(\varrho, \zeta)^\nu \sim \Psi_d(\rho', \zeta)^\nu$.

Proof. From Lemma 5.7 we know that $\rho - \rho' = O(\frac{1}{n})$ and from Lemma 5.8 in connection with Lemma 5.6 we can conclude that $\rho - \varrho = O(\frac{1}{n})$ and thus $\rho' - \varrho = O(\frac{1}{n})$. We use Taylor's formula, which entails that

$$\Psi_d(\varrho + O(\frac{1}{n}), \zeta) - \Psi_d(\varrho, \zeta) = O(\frac{1}{n}) \frac{\partial}{\partial \varrho} \Psi_d(\varrho, \zeta) + O(\frac{1}{n^2}).$$

Deriving Ψ_d we get

$$\begin{aligned} \frac{\partial}{\partial x} \Psi_d(x, \zeta) = & (1-x)^{-d-1} x^{\frac{2x-1}{1-x}} \exp\left(\frac{\zeta}{d} \frac{1-x^d - (1-x)^d}{(1-x)^d}\right) \\ & \cdot \left(c(1-x)(x-x^d) + (1-x)^d x \ln x\right) \end{aligned}$$

which gives $\frac{\partial}{\partial x} \Psi_d(x, \zeta) = 0$ by plugging in (5.2) for $\ln \varrho$ and thus $\Psi_d(\varrho + O(\frac{1}{n}), \zeta) - \Psi_d(\varrho, \zeta) = O(\frac{1}{n^2})$ which together with the fact that $\nu \leq n$ results in $\Psi_d(\rho', \zeta)^\nu \sim \Psi_d(\varrho, \zeta)^\nu$. \square

Proof of Theorem 5.2. Using Lemma 5.9 we know

$$\mathbb{P}[\mathcal{N}(H_d(n, p)) = \nu] \sim \binom{n}{\nu} c_d(\nu, p) (1-p)^{\binom{n}{d} - \binom{\nu}{d} - \binom{n-\nu}{d}}$$

Together with Corollary 5.4 we solve for $c_d(\nu, p)$ and get:

$$c_d(\nu, p) \sim \left(\binom{n}{\nu} (1-p)^{\binom{n}{d} - \binom{\nu}{d} - \binom{n-\nu}{d}} \sqrt{2\pi n \frac{\rho' (1-\rho' + c(d-1)(\rho' - \rho'^{d-1}))}{(1-c(d-1)\rho'^{d-1})^2}} \right)^{-1}$$

Using Stirling's formula we have

$$\binom{n}{\nu} = \binom{n}{(1-\rho')n} \sim \left(\sqrt{2\pi n \rho' (1-\rho')} \rho'^{\rho' n} (1-\rho')^{(1-\rho')n} \right)^{-1}$$

and thus

$$c_d(\nu, p) \sim \rho'^{\rho' n} (1-\rho')^{(1-\rho')n} (1-p)^{\binom{n-\nu}{d} + \binom{\nu}{d} - \binom{n}{d}} \underbrace{\sqrt{\frac{(1-\rho')(1-c(d-1)\rho'^{d-1})^2}{1-\rho'+c(d-1)(\rho'-\rho'^{d-1})}}}_{u_d(\rho', c)}. \quad (5.3)$$

Now we estimate the $(1-p)^{\dots}$ -term. First we focus on $d = 2$:

$$\begin{aligned} (1-p)^{\binom{n-\nu}{2} + \binom{\nu}{2} - \binom{n}{2}} &= (1-p)^{\binom{\rho' n}{2} + \binom{(1-\rho')n}{2} - \binom{n}{2}} \\ &= (1-p)^{-\rho'(1-\rho')n^2} \\ &\sim e^{p\rho'(1-\rho')n^2} e^{p^2\rho'(1-\rho')n^2/2} \\ &\sim e^{c\rho'(1-\rho')(n+1)} e^{c^2\rho'(1-\rho')/2} \end{aligned}$$

Furthermore note that $u_2(\rho', c) = 1 - c\rho'$ in (5.3) and thus

$$\begin{aligned} c_2(\nu, p) &\sim \rho'^{\rho' n} (1-\rho')^{(1-\rho')n} (1-p)^{-\rho'(1-\rho')n^2} (1-c\rho') \\ &\sim \rho'^{\rho' n} (1-\rho')^{(1-\rho')n} e^{c\rho'(1-\rho')(n+1)} e^{c^2\rho'(1-\rho')/2} (1-c\rho') \\ &\sim \rho'^{\rho' n} (1-\rho')^{(1-\rho')n} e^{c\rho'(1-\rho')n} e^{c^2\rho'(1-\rho')/2} (1-c\rho') \end{aligned}$$

Now we use $(1-\rho')n = \nu$ to eliminate n and $c = \zeta(1-\rho')^{-1} \exp\left(\frac{\rho'}{(1-\rho')n} + O(n^{-2})\right)$ (see Lemma 5.8) to replace c by ζ .

$$\begin{aligned} c_2(\nu, p) &\sim \rho'^{\frac{\rho'\nu}{1-\rho'}} (1-\rho')^\nu e^{c\rho'\nu} e^{c^2\rho'(1-\rho')/2} (1-c\rho') \\ &\sim \rho'^{\frac{\rho'\nu}{1-\rho'}} (1-\rho')^\nu e^{\zeta(1-\rho')^{-1}(1+\frac{\rho'}{(1-\rho')n})\rho'\nu} e^{\zeta^2\rho'(1-\rho')^{-1}/2} (1-\zeta\frac{\rho'}{1-\rho'}) \\ &\sim \rho'^{\frac{\rho'\nu}{1-\rho'}} (1-\rho')^\nu e^{\zeta\frac{\rho'\nu}{1-\rho'}} e^{\zeta\frac{\rho'^2}{1-\rho'}} e^{\zeta^2\frac{\rho'}{2(1-\rho')}} (1-\zeta\frac{\rho'}{1-\rho'}) \end{aligned}$$

Using Lemma 5.10 and (5.2) (which gives $\varrho = e^{-\zeta}$) we get:

$$\begin{aligned} c_2(\nu, p) &\sim \varrho^{\frac{\rho'\nu}{1-\varrho}} (1-\varrho)^\nu e^{\zeta\frac{\rho'\nu}{1-\varrho}} e^{\frac{2\zeta\varrho^2+2\zeta\varrho+\zeta^2\varrho}{2(1-\varrho)}} (1-\zeta\frac{\varrho}{1-\varrho}) \\ &= e^{-\zeta\frac{\rho'\nu}{1-\varrho}} (1-\varrho)^\nu e^{\zeta\frac{\rho'\nu}{1-\varrho}} e^{\frac{2\zeta\varrho^2+2\zeta\varrho+\zeta^2\varrho}{2(1-\varrho)}} (1-\zeta\frac{\varrho}{1-\varrho}) \\ &= (1-e^{-\zeta})^\nu e^{\frac{2\zeta\varrho^2+2\zeta\varrho+\zeta^2\varrho}{2(1-\varrho)}} (1-\zeta\frac{\varrho}{1-\varrho}) \\ &= (1-e^{-\zeta})^\nu \exp\left(\frac{2\zeta e^{-2\zeta} + 2\zeta e^{-\zeta} + \zeta^2 e^{-\zeta}}{2(1-e^{-\zeta})}\right) (1-\zeta\frac{e^{-\zeta}}{1-e^{-\zeta}}) \end{aligned}$$

In order to substitute ρ' by ϱ in the calculations above we also made use of the fact that $\rho' - \varrho = O(\frac{1}{n})$ (see the proof of Lemma 5.10)) and that the part of the function c_2 which is independent of n is continuous in ρ' .

Now the same procedure for $d > 2$:

$$\begin{aligned}
 (1-p)^{\binom{\rho'n}{d} + \binom{(1-\rho')n}{d} - \binom{n}{d}} &\sim \exp \left(\left(p + \frac{p^2}{2} \right) \left(\binom{n}{d} - \binom{\rho'n}{d} - \binom{(1-\rho')n}{d} \right) \right) \\
 &\sim \exp \left(\binom{n}{d} p (1 - \rho'^d - (1 - \rho')^d) \right) \\
 &\quad \cdot \exp \left(\binom{n}{d} \frac{p^2}{2} (1 - \rho'^d - (1 - \rho')^d) \right) \\
 &\quad \cdot \exp \left(\binom{n}{d} \binom{d}{2} n^{-1} \left(p + \frac{p^2}{2} \right) ((1 - \rho')\rho'^{d-1} + \rho'(1 - \rho')^{d-1}) \right)
 \end{aligned}$$

For $d > 2$ we have $\binom{n}{d} \frac{p^2}{2} = o(1)$ and thus the second factor is negligible. For the third factor we have

$$\binom{n}{d} \binom{d}{2} n^{-1} \left(p + \frac{p^2}{2} \right) = \frac{cn}{d} \binom{d}{2} n^{-1} \left(1 + \frac{p}{2} \right) \sim \frac{c(d-1)}{2}$$

and thus

$$\begin{aligned}
 (1-p)^{\binom{\rho'n}{d} + \binom{(1-\rho')n}{d} - \binom{n}{d}} &\sim \exp \left(\frac{cn}{d} (1 - \rho'^d - (1 - \rho')^d) \right) \\
 &\quad \cdot \exp \left(\frac{c(d-1)}{2} ((1 - \rho')\rho'^{d-1} + \rho'(1 - \rho')^{d-1}) \right)
 \end{aligned}$$

Now we plug this into (5.3) and reformulate again in terms of $\nu = (1 - \rho')n$ and ζ (using

Lemma 5.8).

$$\begin{aligned}
 c_d(\nu, p) &\sim \rho'^{\rho' n} (1 - \rho')^{(1 - \rho') n} (1 - p)^{\binom{n - \nu}{d} + \binom{\nu}{d} - \binom{n}{d}} u_d(\rho', c) \\
 &\sim \rho'^{\rho' n} (1 - \rho')^{(1 - \rho') n} \exp\left(\frac{cn}{d} (1 - \rho'^d - (1 - \rho')^d)\right) \\
 &\quad \cdot \underbrace{\exp\left(\frac{c(d - 1)}{2} ((1 - \rho') \rho'^{d-1} + \rho' (1 - \rho')^{d-1})\right)}_{\tilde{u}_d(\rho', c)} u_d(\rho', c) \\
 &= \rho'^{\frac{\rho' \nu}{1 - \rho'}} (1 - \rho')^\nu \exp\left(\frac{c\nu}{d} \frac{1 - \rho'^d - (1 - \rho')^d}{1 - \rho'}\right) \tilde{u}_d(\rho', c) \\
 &\sim \rho'^{\frac{\rho' \nu}{1 - \rho'}} (1 - \rho')^\nu \exp\left(\frac{\zeta(1 - \rho')^{1-d} (1 + \frac{\rho'}{(1 - \rho')^n}) \nu}{d} \frac{1 - \rho'^d - (1 - \rho')^d}{1 - \rho'}\right) \\
 &\quad \cdot \tilde{u}_d(\rho', \zeta(1 - \rho')^{1-d}) \\
 &\sim \rho'^{\frac{\rho' \nu}{1 - \rho'}} (1 - \rho')^\nu \exp\left(\frac{\zeta(1 + \frac{\rho'}{\nu}) \nu}{d} \frac{1 - \rho'^d - (1 - \rho')^d}{(1 - \rho')^d}\right) \tilde{u}_d(\rho', \zeta(1 - \rho')^{1-d}) \\
 &\sim \rho'^{\frac{\rho' \nu}{1 - \rho'}} (1 - \rho')^\nu \exp\left(\frac{\zeta \nu}{d} \frac{1 - \rho'^d - (1 - \rho')^d}{(1 - \rho')^d}\right) \\
 &\quad \cdot \underbrace{\exp\left(\zeta \rho' \frac{1 - \rho'^d - (1 - \rho')^d}{d(1 - \rho')^d}\right)}_{u_d^*(\rho', \zeta(1 - \rho')^{1-d})} \tilde{u}_d(\rho', \zeta(1 - \rho')^{1-d})
 \end{aligned}$$

Using Lemma 5.10 and the fact that the constant part is continuous in ρ' we get:

$$\begin{aligned}
 c_d(\nu, p) &\sim \varrho^{\frac{\rho \nu}{1 - \varrho}} (1 - \varrho)^\nu \exp\left(\frac{\zeta \nu}{d} \frac{1 - \varrho^d - (1 - \varrho)^d}{(1 - \varrho)^d}\right) u_d^*(\varrho, \zeta(1 - \varrho)^{1-d}) \\
 &= \Psi_d(\zeta)^\nu u_d^*(\varrho, \zeta(1 - \varrho)^{1-d})
 \end{aligned}$$

where

$$\begin{aligned}
 u_d^*(\varrho, \zeta(1-\varrho)^{1-d}) &= \exp\left(\zeta\varrho\frac{1-\varrho^d-(1-\varrho)^d}{d(1-\varrho)^d}\right) \\
 &\quad \cdot \exp\left(\frac{\zeta(1-\varrho)^{1-d}(d-1)}{2}((1-\varrho)\varrho^{d-1} + \varrho(1-\varrho)^{d-1})\right) \\
 &\quad \cdot \sqrt{\frac{(1-\varrho)(1-\zeta(1-\varrho)^{1-d}(d-1)\varrho^{d-1})^2}{1-\varrho + \zeta(1-\varrho)^{1-d}(d-1)(\varrho - \varrho^{d-1})}} \\
 &= \exp\left(\zeta\varrho\frac{1-\varrho^d-(1-\varrho)^d}{d(1-\varrho)^d} + \frac{\zeta(d-1)\varrho}{2}\left(\left(\frac{\varrho}{1-\varrho}\right)^{d-2} + 1\right)\right) \\
 &\quad \cdot \frac{1 - \zeta(d-1)\left(\frac{\varrho}{1-\varrho}\right)^{d-1}}{\sqrt{1 + \zeta(d-1)\frac{\varrho - \varrho^{d-1}}{(1-\varrho)^d}}}
 \end{aligned}$$

□

The reason why our result for $d = 2$ differs from the one given by Stepanov [1970]

$$c_2(n, t) \sim \left(1 - \frac{nt}{e^{nt} - 1}\right)(1 - e^{-nt})^n$$

is twofold. First we have $p = 1 - e^{-t}$ instead of $p = t$ and second, in contrast to the general notion, we do not have $pn = c$ but rather $p(n-1) = c$.

5.4 Connectivity Probability and the Number of Connected Graphs

The proof is similar to that of Theorem 1 in Coja-Oghlan et al. [2006]. First we give a

Lemma 5.11.

$$\mathbb{P}[\mathcal{N}(H_d(n, p)) = \nu \wedge \mathcal{M}(H_d(n, p)) = \mu] \sim \frac{c_d(\nu, \mu)}{2\pi\sqrt{\rho(1-\rho)(1-\rho^d)n\frac{cn}{d}w_d(c, n)}}$$

where

$$\begin{aligned}
 w_d(c, n) &= \exp\left(\frac{c(d-1)(\rho + \rho^{d-1} - 2\rho^d)}{2(1-\rho)} + \frac{(1-\rho^d)c^2n}{2d\binom{n-1}{d-1}} \cdot \frac{1-\rho^d-(1-\rho)^d}{(1-\rho)^d}\right) \\
 &\quad \cdot \left(\rho^{\frac{\rho}{1-\rho}}(1-\rho)\right)^\nu \left(\frac{1-\rho^d}{(1-\rho)^d}\right)^\mu
 \end{aligned}$$

Proof. Use the statements in the proof of Theorem 1 in Coja-Oghlan et al. [2006], especially Lemma 10 which gives Equation (15) in Coja-Oghlan et al. [2006]. □

Proof of Theorem 5.1. Given numbers ν, μ such that $\zeta = d\mu/\nu$ satisfies $c_0 \leq \zeta \leq C_0$ for some constants $C_0 > c_0 > \frac{d}{d-1}$, our goal is to compute the probability that $H_d(\nu, \mu)$ is connected. We reduce this problem to the problem of computing the probability that the largest component of a random hypergraph $H_d(n, p)$ has order exactly ν and size exactly μ , where the edge probability $p = c/\binom{n-1}{d-1}$ is chosen appropriately.

Let us first specify the edge probability p . By Corollary 5.4 for each $c_m > \frac{1}{d-1}$ there is an $0 < \rho_m = \rho_m(c_m) < 1$ such that $\rho_m = \exp(c_m(\rho_m^{d-1} - 1))$.

By [Coja-Oghlan et al., 2006, Lemma 7] we can choose c_m so that $\zeta = (1 - \rho_m^d)c_m/(1 - \rho_m)$. In addition, we choose n to be the largest integer such that $(1 - \rho_m)n \leq \nu$, and we set $p_m = c_m \binom{n-1}{d-1}^{-1}$. Then $\nu - 1 \leq (1 - \rho_m)n \leq \nu$, because $0 < \rho_m < 1$. Moreover, using the relation $\zeta = (1 - \rho_m^d)c_m/(1 - \rho_m)$, we see that ρ_m satisfies the equation

$$\rho_m = \exp\left(c_m(\rho_m^{d-1} - 1)\right) = \exp\left(-\zeta(1 - \rho_m) \cdot \frac{1 - \rho_m^{d-1}}{1 - \rho_m^d}\right) \quad (\text{cf. (5.1)}). \quad (5.4)$$

While $\nu = \lceil (1 - \rho_m)n \rceil$, it is more convenient to work with an edge probability $p = c \binom{n-1}{d-1}^{-1}$ such that $\nu = (1 - \rho)n$, where ρ is the solution to the equation $\rho = \exp(c(\rho^{d-1} - 1))$ (cf. Corollary 5.5).

Using Lemma 5.11 we know that

$$\mathbb{P}[\mathcal{N}(H_d(n, p)) = \nu \wedge \mathcal{M}(H_d(n, p)) = \mu] = \frac{c_d(\nu, \mu)}{2\pi \sqrt{\rho(1 - \rho)(1 - \rho^d)n^{\frac{cn}{d}} w_d(c, n)}}$$

where

$$\begin{aligned} w_d(c, n) \sim \exp\left(\frac{c(d-1)(\rho + \rho^{d-1} - 2\rho^d)}{2(1 - \rho)} + \frac{(1 - \rho^d)c^2 n}{2d \binom{n-1}{d-1}} \cdot \frac{1 - \rho^d - (1 - \rho)^d}{(1 - \rho)^d}\right) \\ \cdot \left(\rho^{\frac{\rho}{1-\rho}}(1 - \rho)\right)^\nu \left(\frac{1 - \rho^d}{(1 - \rho)^d}\right)^\mu \end{aligned}$$

Since $\nu = (1 - \rho)n$ and $\mu = (1 - \rho^d)\frac{cn}{d}$ we can plug in Corollary 5.5 and get

$$\begin{aligned} c_d(\nu, \mu)^2 &= \frac{4\pi^2 \rho(1 - \rho)(1 - \rho^d)n^{\frac{cn}{d}} w_d(c, n)^2}{4\pi^2 n^2} \\ &\quad \cdot \frac{(1 - c(d-1)\rho^{d-1})^2}{\frac{c}{d}\rho(1 - \rho + c(d-1)(\rho - \rho^{d-1}))(1 - \rho^d) - c^2\rho^2(1 - \rho^{d-1})^2} \\ &= \frac{(1 - \rho)(1 - \rho^d)(1 - c(d-1)\rho^{d-1})^2}{(1 - \rho + c(d-1)(\rho - \rho^{d-1}))(1 - \rho^d) - cd\rho(1 - \rho^{d-1})^2} w_d(c, n)^2 \end{aligned}$$

Now we want to reformulate this in terms of $\zeta = \frac{d\mu}{\nu} = \frac{(1 - \rho^d)c}{1 - \rho}$ and $\nu = (1 - \rho)n$ instead

of c and n . Let $\omega_d(\zeta, \nu) := w_d\left(\frac{(1-\rho)\zeta}{1-\rho^d}, \frac{\nu}{1-\rho}\right)$

$$\begin{aligned}
 c_d(\nu, \mu)^2 &= \frac{(1-\rho)(1-\rho^d)(1-\frac{(1-\rho)\zeta}{1-\rho^d}(d-1)\rho^{d-1})^2}{(1-\rho+\frac{(1-\rho)\zeta}{1-\rho^d}(d-1)(\rho-\rho^{d-1}))(1-\rho^d)-\frac{(1-\rho)\zeta}{1-\rho^d}d\rho(1-\rho^{d-1})^2}\omega_d(\zeta, \nu)^2 \\
 &= \frac{(1-\rho^d)(1-\frac{(1-\rho)\zeta}{1-\rho^d}(d-1)\rho^{d-1})^2}{(1+\frac{\zeta}{1-\rho^d}(d-1)(\rho-\rho^{d-1}))(1-\rho^d)-\frac{\zeta}{1-\rho^d}d\rho(1-\rho^{d-1})^2}\omega_d(\zeta, \nu)^2 \\
 &= \frac{\frac{1-\rho^d}{(1-\rho^d)^2}(1-\rho^d-(1-\rho)\zeta(d-1)\rho^{d-1})^2}{(1-\rho^d+\zeta(d-1)(\rho-\rho^{d-1}))- \frac{\zeta}{1-\rho^d}d\rho(1-\rho^{d-1})^2}\omega_d(\zeta, \nu)^2 \\
 &= \frac{(1-\rho^d-(1-\rho)\zeta(d-1)\rho^{d-1})^2}{(1-\rho^d+\zeta(d-1)(\rho-\rho^{d-1}))(1-\rho^d)-\zeta d\rho(1-\rho^{d-1})^2}\omega_d(\zeta, \nu)^2
 \end{aligned}$$

which is for $d = 2$:

$$\begin{aligned}
 c_2(\nu, \mu)^2 &= \frac{(1-\rho^2-(1-\rho)\zeta\rho)^2}{(1-\rho^2)(1-\rho^2)-2\zeta\rho(1-\rho)^2}\omega_2(\zeta, \nu)^2 \\
 &= \frac{(1-\rho)^2(1+\rho-\zeta\rho)^2}{(1-\rho)^2(1+\rho)^2-2\zeta\rho(1-\rho)^2}\omega_2(\zeta, \nu)^2 \\
 &= \frac{(1+\rho-\zeta\rho)^2}{(1+\rho)^2-2\zeta\rho}\omega_2(\zeta, \nu)^2.
 \end{aligned}$$

with

$$\begin{aligned}
 \omega_2(\zeta, \nu) &\sim \exp\left(\frac{\frac{(1-\rho)\zeta}{1-\rho^2}(2\rho-2\rho^2)}{2(1-\rho)} + \frac{(1-\rho^2)(\frac{(1-\rho)\zeta}{1-\rho^2})^2}{4} \cdot \frac{1-\rho^2-(1-\rho)^2}{(1-\rho)^2}\right) \\
 &\quad \left(\rho^\rho(1-\rho)^{1-\rho}\left(\frac{1-\rho^2}{(1-\rho)^2}\right)^{\frac{(1-\rho)\zeta(1-\rho^2)}{2(1-\rho^2)}}\right)^{\frac{\nu}{1-\rho}} \\
 &= \exp\left(\frac{\zeta\rho(1-\rho)}{1-\rho^2} + \frac{\zeta^2(1-\rho^2-(1-\rho)^2)}{4(1-\rho^2)}\right) \cdot \left(\rho^{\frac{\rho}{1-\rho}}(1-\rho)\left(\frac{1+\rho}{1-\rho}\right)^{\frac{\zeta}{2}}\right)^\nu \\
 &= \exp\left(\frac{\zeta\rho}{1+\rho} + \frac{\zeta^2(2\rho-2\rho^2)}{4(1-\rho^2)}\right) \cdot \left(\rho^{\frac{\rho}{1-\rho}}(1-\rho)\left(\frac{1+\rho}{1-\rho}\right)^{\frac{\zeta}{2}}\right)^\nu \\
 &= \exp\left(\frac{\zeta\rho}{1+\rho} + \frac{\zeta^2\rho(1-\rho)}{2(1-\rho^2)}\right) \cdot \left(\rho^{\frac{\rho}{1-\rho}}(1-\rho)\left(\frac{1+\rho}{1-\rho}\right)^{\frac{\zeta}{2}}\right)^\nu \\
 &= \exp\left(\frac{2\zeta\rho+\zeta^2\rho}{2(1+\rho)}\right) \cdot \left(\rho^{\frac{\rho}{1-\rho}}(1-\rho)\left(\frac{1+\rho}{1-\rho}\right)^{\frac{\zeta}{2}}\right)^\nu
 \end{aligned}$$

and thus

$$c_2(\nu, \mu) \sim \frac{1 + \rho - \zeta\rho}{\sqrt{(1 + \rho)^2 - 2\zeta\rho}} \exp\left(\frac{2\zeta\rho + \zeta^2\rho}{2(1 + \rho)}\right) \cdot \Phi_2(\rho, \zeta)^\nu.$$

Now for the case $d > 2$:

$$\begin{aligned} \omega_d(\zeta, \nu) &\sim \exp\left(\frac{\frac{(1-\rho)\zeta}{1-\rho^d}(d-1)(\rho + \rho^{d-1} - 2\rho^d)}{2(1-\rho)}\right) \cdot \left(\rho^\rho(1-\rho)^{1-\rho}\left(\frac{1-\rho^d}{(1-\rho)^d}\right)^{\frac{(1-\rho)\zeta(1-\rho^d)}{(1-\rho^d)d}}\right)^{\frac{\nu}{1-\rho}} \\ &= \exp\left(\frac{\zeta(d-1)(\rho + \rho^{d-1} - 2\rho^d)}{2(1-\rho^d)}\right) \cdot \left(\rho^{\frac{\rho}{1-\rho}}(1-\rho)\left(\frac{1-\rho^d}{(1-\rho)^d}\right)^{\frac{\zeta}{d}}\right)^\nu, \end{aligned}$$

thus

$$\begin{aligned} c_d(\nu, \mu) &\sim \frac{1 - \rho^d - (1 - \rho)\zeta(d-1)\rho^{d-1}}{\sqrt{(1 - \rho^d + \zeta(d-1)(\rho - \rho^{d-1}))(1 - \rho^d) - \zeta d\rho(1 - \rho^{d-1})^2}} \\ &\quad \cdot \exp\left(\frac{\zeta(d-1)(\rho - 2\rho^d + \rho^{d-1})}{2(1 - \rho^d)}\right) \cdot \Phi_d(\rho, \zeta)^\nu. \end{aligned}$$

It was already shown in [Coja-Oghlan et al., 2006, Lemma 12] that $\Phi_d(\rho, \zeta)^\nu \sim \Phi_d(\rho_m, \zeta)^\nu$. Together with the fact that the constant part of c_d is continuous in ρ and that $\rho - \rho_m = O(\frac{1}{n})$ the assertion follows. \square

Note that this result fits the result of Bender et al. [1990] as we will show next. Their result is:

$$\begin{aligned} c_2(\nu, \mu) &\sim \exp\left(x(x+1)(1-y) + \log(1-x+xy) - \frac{\log(1-x+xy^2)}{2}\right) \cdot \left(\frac{2e^{-x}y^{1-x}}{\sqrt{1-y^2}}\right)^\nu \\ &= \frac{1-x+xy}{\sqrt{1-x+xy^2}} \exp(x(x+1)(1-y)) \cdot \left(\frac{2e^{-x}y^{1-x}}{\sqrt{1-y^2}}\right)^\nu \end{aligned}$$

with $x = \frac{\zeta}{2}$ and $y = \frac{1-\rho}{1+\rho}$.

$$\begin{aligned}
 c_2(\nu, \mu) &\sim \frac{1 - \frac{\zeta}{2} + \frac{\zeta}{2} \frac{1-\rho}{1+\rho}}{\sqrt{1 - \frac{\zeta}{2} + \frac{\zeta}{2} \left(\frac{1-\rho}{1+\rho}\right)^2}} \exp\left(\frac{\zeta}{2} \left(\frac{\zeta}{2} + 1\right) \left(1 - \frac{1-\rho}{1+\rho}\right)\right) \cdot \left(\frac{2e^{-\frac{\zeta}{2}} \left(\frac{1-\rho}{1+\rho}\right)^{1-\frac{\zeta}{2}}}{\sqrt{1 - \left(\frac{1-\rho}{1+\rho}\right)^2}}\right)^\nu \\
 &= \frac{1 + \rho - \zeta\rho}{\sqrt{(1+\rho)^2 - 2\zeta\rho}} \exp\left(\frac{\zeta(\frac{\zeta}{2} + 1)\rho}{1 + \rho}\right) \cdot \left(\frac{2\rho^{\frac{1+\rho}{2(1-\rho)}} \left(\frac{1+\rho}{1-\rho}\right)^{\frac{\zeta}{2}} (1-\rho)}{2\sqrt{\rho}}\right)^\nu \\
 &= \frac{1 + \rho - \zeta\rho}{\sqrt{(1+\rho)^2 - 2\zeta\rho}} \exp\left(\frac{\zeta^2\rho + 2\zeta\rho}{2(1+\rho)}\right) \cdot \left(\rho^{\frac{1+\rho}{2(1-\rho)} - \frac{1}{2}} \left(\frac{1+\rho}{1-\rho}\right)^{\frac{\zeta}{2}} (1-\rho)\right)^\nu \\
 &= \frac{1 + \rho - \zeta\rho}{\sqrt{(1+\rho)^2 - 2\zeta\rho}} \exp\left(\frac{\zeta^2\rho + 2\zeta\rho}{2(1+\rho)}\right) \cdot \left(\rho^{\frac{1+\rho-1+\rho}{2(1-\rho)}} \left(\frac{1+\rho}{1-\rho}\right)^{\frac{\zeta}{2}} (1-\rho)\right)^\nu \\
 &= \frac{1 + \rho - \zeta\rho}{\sqrt{(1+\rho)^2 - 2\zeta\rho}} \exp\left(\frac{\zeta^2\rho + 2\zeta\rho}{2(1+\rho)}\right) \cdot \left(\rho^{\frac{\rho}{1-\rho}} \left(\frac{1+\rho}{1-\rho}\right)^{\frac{\zeta}{2}} (1-\rho)\right)^\nu
 \end{aligned}$$

5.5 Edge Distribution of Connected Hypergraphs

Proof of Theorem 5.3. We want to calculate the distribution of the number of edges for a connected graph $H_d(\nu, p)$ via finding the edge distribution for the giant component of $H_d(n, p)$ where n is chosen such that the giant component of $H_d(n, p)$ has about ν vertices. As in the proof of Theorem 5.2 we choose n to be the largest integer such that $(1-\rho)n \leq \nu$ where ρ is the solution to $\rho = \exp(p \binom{n-1}{d-1} (\rho^{d-1} - 1))$. Lemma 5.7 gives that $\nu - (1-\rho)n = O(1)$, thus we can directly apply Corollaries 5.4 and 5.5.

It was already shown in [Coja-Oghlan et al., 2006, Theorem 3] that $\mu_g = \frac{\zeta\nu}{d(1-\rho)^d} (1 - \rho^d) \sim \frac{cn}{d} (1 - \rho^d)$.

$$\begin{aligned}
 g(c, y) &:= \mathbb{P}[|E(H_d(\nu, p))| = \mu_g + y \mid H_d(\nu, p) \text{ is connected}] \\
 &= \mathbb{P}[\mathcal{M}(H_d(n, p)) = \mu_g + y \mid \mathcal{N}(H_d(n, p)) = \nu] \\
 &\sim \frac{\mathbb{P}[\mathcal{M}(H_d(n, p)) = \mu_g + y \wedge \mathcal{N}(H_d(n, p)) = (1 - \rho)n]}{\mathbb{P}[\mathcal{N}(H_d(n, p)) = (1 - \rho)n]} \\
 &= \frac{1}{2\pi n} \frac{1 - c(d-1)\rho^{d-1}}{\sqrt{\frac{c}{d}\rho(1 - \rho + c(d-1)(\rho - \rho^{d-1}))}(1 - \rho^d) - c^2\rho^2(1 - \rho^{d-1})^2}} \\
 &\quad \cdot \exp\left(-\frac{y^2 d}{2cn}\left(1 - \frac{cd\rho(1 - \rho^{d-1})^2}{1 - \rho + c(d-1)(\rho - \rho^{d-1})} - \rho^d\right)^{-1}\right) \\
 &\quad \cdot \left(2\pi n \frac{\rho(1 - \rho + c(d-1)(\rho - \rho^{d-1}))}{(1 - c(d-1)\rho^{d-1})^2}\right)^{1/2} \\
 &= \frac{1}{\sqrt{2\pi n}} \left(\frac{c}{d}(1 - \rho^d) - \frac{c^2\rho^2(1 - \rho^{d-1})^2}{\rho(1 - \rho + c(d-1)(\rho - \rho^{d-1}))}\right)^{-1/2} \\
 &\quad \cdot \exp\left(-\frac{y^2 d}{2cn}\left(1 - \frac{cd\rho(1 - \rho^{d-1})^2}{1 - \rho + c(d-1)(\rho - \rho^{d-1})} - \rho^d\right)^{-1}\right) \\
 &= \frac{1}{\sqrt{2\pi}\sigma_g} \exp\left(-\frac{y^2}{2\sigma_g^2}\right)
 \end{aligned}$$

with

$$\sigma_g^2 := \frac{cn}{d}\left(1 - \frac{cd\rho(1 - \rho^{d-1})^2}{1 - \rho + c(d-1)(\rho - \rho^{d-1})} - \rho^d\right)$$

Now we want to reformulate this in terms of $\nu := (1 - \rho)n$ and ζ chosen such that $\zeta/\binom{\nu-1}{d-1} = c/\binom{n-1}{d-1} = p$. This means nothing but $\zeta \sim c(1 - \rho)^{d-1}$ and thus

$$\begin{aligned}
 \sigma_g^2 &= \frac{\zeta\nu}{d(1 - \rho)^d}\left(1 - \frac{\frac{\zeta}{(1-\rho)^{d-1}}d\rho(1 - \rho^{d-1})^2}{1 - \rho + \frac{\zeta}{(1-\rho)^{d-1}}(d-1)(\rho - \rho^{d-1})} - \rho^d\right) \\
 &= \frac{\zeta\nu}{d(1 - \rho)^d}\left(1 - \frac{\zeta d\rho(1 - \rho^{d-1})^2}{(1 - \rho)^d + \zeta(d-1)(\rho - \rho^{d-1})} - \rho^d\right)
 \end{aligned}$$

and

$$\mu_g = \frac{\zeta\nu}{d(1 - \rho)^d}(1 - \rho^d)$$

For $d = 2$ we have $\rho = e^{-\zeta}$ and thus

$$\begin{aligned}
 \sigma_g^2 &= \frac{\zeta\nu}{2(1 - \rho)^2}\left(1 - \frac{2\zeta\rho(1 - \rho)^2}{(1 - \rho)^2} - \rho^2\right) \\
 &= \frac{\zeta\nu}{2(1 - e^{-\zeta})^2}(1 - 2\zeta e^{-\zeta} - e^{-2\zeta})
 \end{aligned}$$

and

$$\begin{aligned}\mu_g &= \frac{\zeta\nu}{2(1-\rho)^2}(1-\rho^2) \\ &= \frac{\zeta\nu(1+\rho)}{2(1-\rho)} \\ &= \frac{\zeta\nu(1+e^{-\zeta})}{2(1-e^{-\zeta})}\end{aligned}$$

□

Part II

Random Intersection Graphs

Chapter 6

Introduction

Trying to model the component evolution of the molecule network from Frömmel et al. [2003] with the standard random graph model introduced by Erdős and Rényi [1960] which was covered in Part I leads to the disappointing result shown in Figure 6.1.

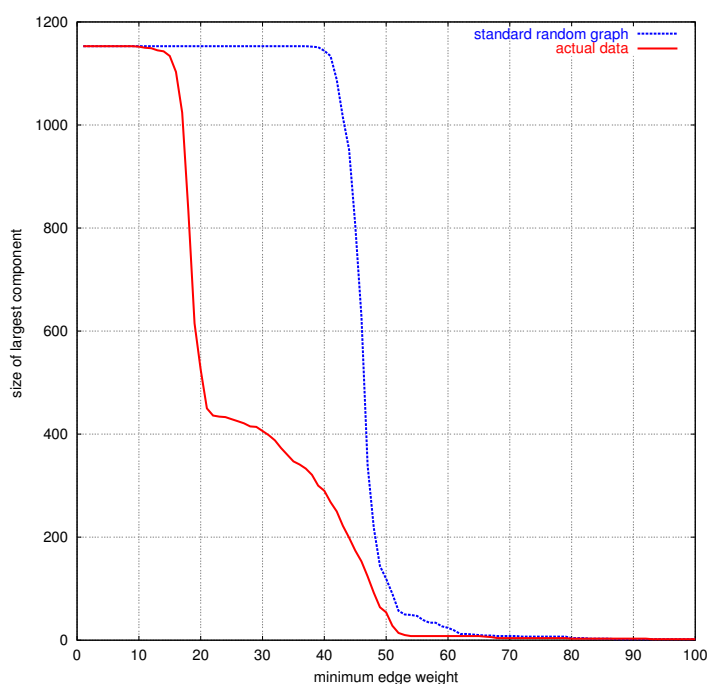


Figure 6.1: Largest component in the protein graph versus the Erdős-Rényi-model.

The largest component growing at significantly slow speed between the edge weights 20 and 50 in the experimental network behaved completely differently in the model. In accordance with the theoretical results the largest component in the model jumps almost at once from only a few vertices to the almost complete graph.

The reasons for this mismatch and an attempt for developing a model solving these

issues are the topics of this part.

6.1 A Different Model for Random Graphs

The random graph model by Erdős and Rényi [1960] (denoted by $G_{n,p}$) considers a fixed set of n vertices and edges that exist with a certain probability $p = p(n)$, independently from each other. In addition to the difference in component behaviour mentioned above, it also lacks many of the commonly observed properties of real-world networks (e.g. scale free degree distribution and clustering, see for instance Albert and Barabási [2002]). One of the underlying reasons that are responsible for this mismatch is precisely the independence of the edges, in other words the missing transitivity. In a real-world network, relations between vertices x and y on the one hand and between vertices y and z on the other hand suggest a connection of some sort between vertices x and z .

6.1.1 Intersection Graphs

An *intersection graph* is a graph $G = (V, E)$ together with a so-called *universal feature set* W . Every vertex $x \in V$ has an assigned *feature set* $W_x \subseteq W$, and the characteristic property of an intersection graph is that two vertices $x, y \in V$ are connected by an edge in E if and only if their feature sets have non-empty intersection:

$$\{x, y\} \in E \Leftrightarrow W_x \cap W_y \neq \emptyset.$$

We call the elements of W *features*. If the feature $w \in W$ is contained in W_x and W_y and thus forces the edge $\{x, y\}$, we say that $\{x, y\}$ is *induced* by w . Furthermore the set of vertices V_w holding a specified feature w (which obviously forms a clique) is called a *feature clique*. Trivially

$$v \in V_w \Leftrightarrow w \in W_v,$$

in which case we say that v and w *see* each other or v *has* w .

As usual, $\Gamma(v)$ denotes the set of neighbours of v , i.e. the set of vertices in V that have features with v in common.

Well studied examples for intersection graphs are interval graphs on the real line (see e.g. Scheinerman [1988]). We will, however, only consider finite sets. Obviously every graph is an intersection graph (simply pick an individual feature assigned only to the two vertices of every edge), but the fewer features we have, the more apparent becomes the structure of the shared features inside the graph.

6.1.2 Random Intersection Graphs

A *random intersection graph* on n vertices with a universal feature set W of size m is a random graph with vertex set $[n]$ where each vertex gets assigned a random set of features by choosing each feature independently with probability p . A sample of this probability space is denoted by $G_{n,m,p}$. We consider now and in the following $m := n^\alpha$, and will usually distinguish two cases: $\alpha > 1$ and $0 < \alpha < 1$. If the probability of $G_{n,m,p}$ having

a property \mathcal{A} tends to 1 with n tending to infinity, we say that $G_{n,m,p}$ has property \mathcal{A} *asymptotically almost surely* (a.a.s.).

A few simple observations. Obviously $G_{n,m,p}$ does exhibit some kind of transitivity: if the edges $\{x, y\}$ and $\{y, z\}$ are induced by the same feature w , then this will also induce the edge $\{x, z\}$. The smaller m is, the ‘simpler’ will be $G_{n,m,p}$, because relatively few cliques will dominate its structure. In the following we will consider the case $m := n^\alpha$. It was shown in Fill et al. [2000] that for $\alpha > 6$ the random intersection graph $G_{n,m,p}$ behaves in many ways like the classical random graph $G_{n,p'}$ with $p' = 1 - (1 - p^2)^m$. We will focus mainly on the case where $0 < \alpha < 1$.

It is sometimes convenient to view the random intersection graph as a random bipartite graph with bipartition (V, W) and random edges between the V and W occurring independently with probability p . A sample from this space will be denoted by $B_{n,m,p}$. Given the bipartite graph, say B , the intersection graph is obtained as $G = B^2[V]$, where we write B^2 for the so-called square of B (where two vertices are connected if their distance is at most 2 in B). B is called a *generator* of G .

6.1.3 Related Work

The model of a random intersection graph $G_{n,m,p}$ has been studied with respect to subgraph appearance in Karoński et al. [1999] and with respect to equivalence to $G_{n,p}$ in Fill et al. [2000] (see also Singer [1995]). Stark has investigated the vertex degree distribution in Stark [2004]. The first two results and some results concerning connectivity and cliques can also be found in Singer [1995].

Extensions to the model have been proposed in Godehardt and Jaworski [2001], who modify the distribution of the sizes of the feature cliques. The practical relevance of the model has been discussed in Newman et al. [2001] and in Guillaume and Latapy [2004].

Studies on the extended model concerning degree distribution were performed by Jaworski et al. [2006] and concerning evolution of the largest component by Rybarczyk [2006]. For related work concerning the different problems studied in the individual chapters also see the notes at the beginning of each chapter.

6.1.4 Overview

In this part we will give first some auxiliary lemmas and then study in three chapters some selected problems on random intersection graphs. The first one (Chapter 7) is the evolution of the giant component, which motivated us to look at this model. The second one is the construction of a minimum clique cover (Chapter 8) which somehow restores the intersection graph structure of a given graph and last but not least we look at the colouring problem (Chapter 9) as a prominent optimization problem which also gives insight into clique (and thus cluster) structure of the graph.

After the theoretical results we present some experimental studies on real-world networks and close with an outlook on open problems on random intersection graphs in connection with complex networks.

6.2 Auxiliary lemmas

The following estimates are used without proof:

$$\binom{a}{b} \leq \left(\frac{ea}{b}\right)^b \quad (6.1)$$

$$\binom{a}{b} \leq a^b \quad (6.2)$$

$$(1-a)^b = (1+o(1))(1-ab) \quad \text{for } 0 < a < 1, ab \rightarrow 0 \quad (6.3)$$

$$e^{-2a} \leq 1-a \leq e^{-a} \quad \text{for } 0 \leq a \leq \frac{1}{2} \quad (6.4)$$

Let X be a non-negative random variable with expectation $\mu := \mathbb{E}[X]$ and variance $\text{Var}[X]$. As a special case of Markov's inequality the first moment method states that

$$\mathbb{P}[X \geq 1] \leq \mu. \quad (6.5)$$

and the second moment method (a special case of Tschebyscheff's inequality) that

$$\mathbb{P}[X = 0] \leq \text{Var}[X]/\mu^2 = \frac{\mathbb{E}[X^2]}{\mu^2} - 1. \quad (6.6)$$

If X is a binomially distributed random variable (n trials, each with probability p), then $\mu = np$ and we shall use the following variants of Chernoff's inequality (see Section 2 in Janson et al. [2000]):

$$\mathbb{P}[X \geq \mu + t] \leq \exp\left(-\frac{t^2}{2(\mu + t/3)}\right) \quad \text{for } t \geq 0, \quad (6.7)$$

$$\mathbb{P}[X \leq \mu - t] \leq \exp\left(-\frac{t^2}{2\mu}\right) \quad \text{for } t \geq 0, \quad (6.8)$$

$$\mathbb{P}[X \geq t] \leq \exp(-t) \quad \text{for } t \geq 7\mu. \quad (6.9)$$

Let $G_{n,m,p}$ be a random intersection graph. We first show that the probability that there is a feature clique which deviates significantly from its expected size is exponentially small.

Lemma 6.1. *Let $X_w := |V_w|$ be the random variable counting the number of vertices of a fixed feature w in a random intersection graph $G_{n,m,p}$ with $m := n^\alpha$ and $\alpha < 1$. Then*

$$\mathbb{P}\left[\exists w \in W : |X_w - pn| > (pn)^{\frac{3}{4}}\right] \leq m \exp\left(-\frac{(pn)^{\frac{1}{2}}}{3}\right).$$

Proof. The number of vertices chosen by a feature is a binomially distributed variable. Its deviation from its expected value can therefore be bounded by Chernoff inequalities (6.7) and (6.8). First let w be fixed:

$$\mathbb{P}\left[X_w > pn + (pn)^{\frac{3}{4}}\right] \leq \exp\left(-\frac{(pn)^{\frac{3}{2}}}{2(pn + (pn)^{\frac{3}{4}}/3)}\right) \leq \frac{1}{2} \exp\left(-\frac{(pn)^{\frac{1}{2}}}{3}\right)$$

$$\mathbb{P} \left[X_w < pn - (pn)^{\frac{3}{4}} \right] \leq \exp \left(-\frac{(pn)^{\frac{3}{2}}}{2pn} \right) \leq \frac{1}{2} \exp \left(-\frac{(pn)^{\frac{1}{2}}}{3} \right).$$

By linearity of expectation (summing over all possible w) and Markov's inequality this implies that

$$\mathbb{P} \left[\exists w \in W : |X_w - pn| > (pn)^{\frac{3}{4}} \right] \leq m \exp \left(-\frac{(pn)^{\frac{1}{2}}}{3} \right).$$

□

Since we are mostly interested in small feature sets, we need only an upper bound on their size.

Lemma 6.2. *Let $X_v := |W_v|$ be the random variable counting the number of features for a fixed vertex v in a random intersection graph $G_{n,m,p}$ with $m := n^\alpha$ and $\alpha < 1$. Then*

$$\mathbb{P} [\exists v \in V : X_v > 2pm] \leq ne^{-\frac{3pm}{8}},$$

and for $pm \leq 3 \ln n$

$$\mathbb{P} [\exists v \in V : X_v > 21 \ln n] \leq \frac{1}{n^{20}}.$$

Proof. Very similarly to the previous lemma, we have for a fixed vertex v

$$\mathbb{P} [X_v > pm + pm] \stackrel{(6.7)}{\leq} \exp \left(-\frac{(pm)^2}{2(pm + pm/3)} \right) = e^{-\frac{3pm}{8}}$$

and for $pm \leq 3 \ln n$

$$\mathbb{P} [X_v > 21 \ln n] \stackrel{(6.9)}{\leq} \exp(-21 \ln n) = \frac{1}{n^{21}}.$$

Again summing over all vertices v yields the statement of the lemma. □

Denote by B the event that none of the events in Lemmas 6.1 and 6.2 occur. In other words, for no $w \in W : |X_w - pn| > \frac{pm}{2}$ and for no $v \in V : X_v > 2pm$ or $X_v > 21 \ln n$. The above lemmas show that (under certain conditions on n , m and p) we have $\mathbb{P} [\bar{B}] \rightarrow 0$. In the following we will often observe that these conditions are indeed satisfied, and then attempt to compute the probability for some other event A . As

$$\mathbb{P} [A] = \mathbb{P} [A|B] \mathbb{P} [B] + \mathbb{P} [A|\bar{B}] \mathbb{P} [\bar{B}] \leq \mathbb{P} [A|B] + \mathbb{P} [\bar{B}],$$

we can then restrict our attention to proving that $\mathbb{P} [A|B] \rightarrow 0$.

Chapter 7

Component Evolution

7.1 Results

The aim of this chapter is to study the evolution of the largest component in the random intersection graph model. Since components are natural candidates for clusters in graphs it is straightforward to analyse their growth in our random model, thereby getting insight into structural peculiarities of the real-world networks. The component structure for $G_{n,p}$ has already been studied in Erdős and Rényi [1960] and there are also results for some models for real-world networks by Chung and Lu [2002] and Bollobás and Riordan [2005].

This chapter is organised as follows. In the next section we describe our results and compare it with the growth of the giant component in $G_{n,p}$. Section 7.2 states some results on branching processes which will be used for the proofs of the results in Section 7.3 and 7.4.

Let $\mathcal{N}(G)$ denote the order (number of vertices) of the largest component of G . Our main theorem is:

Theorem 7.1. *Let $G_{n,m,p}$ be a random intersection graph with $m := n^\alpha$ and $p^2m = \frac{c}{n}$. Furthermore let ρ be the single solution to $\rho = \exp(c(\rho - 1))$ in the interval $(0, 1)$ for $c > 1$. Then we have a.a.s.*

$$\mathcal{N}(G_{n,m,p}) \leq \frac{9}{(1-c)^2} \ln n \quad \text{for } \alpha > 1 \text{ and } c < 1 \quad (7.1)$$

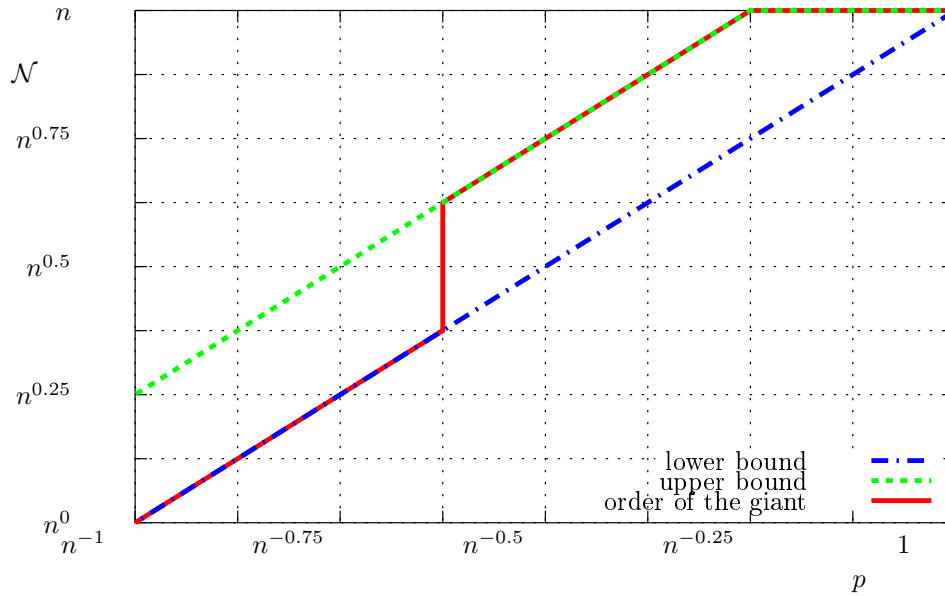
$$\mathcal{N}(G_{n,m,p}) = (1 + o(1))(1 - \rho)n \quad \text{for } \alpha > 1 \text{ and } c > 1 \quad (7.2)$$

$$\mathcal{N}(G_{n,m,p}) \leq \frac{10\sqrt{c}}{(1-c)^2} \sqrt{\frac{n}{m}} \ln m \quad \text{for } \alpha < 1 \text{ and } c < 1 \quad (7.3)$$

$$\mathcal{N}(G_{n,m,p}) = (1 + o(1))(1 - \rho)\sqrt{cmn} \quad \text{for } \alpha < 1 \text{ and } c > 1 \quad (7.4)$$

Furthermore we can prove that the order of the largest component for $\alpha < \frac{1}{2}$ and p small enough is approximately that of a single feature clique, see Section 7.4.1 for details.

As already proven in Singer [1995] the “edge probability” p' (meaning the ratio between present edges and all possible edges) in the random intersection graph is closely concentrated around p^2m . Thus the two results above show that for $\alpha > 1$ the largest


 Figure 7.1: Evolution of the largest component for $\alpha = 0.25$.

component in the intersection graph exhibits a jump from logarithmic order to linear order at $p' = \frac{1}{n}$ which is similar to the $G_{n,p'}$ behaviour. This is also the moment at which in both models the expected degree of a vertex gets larger than 1.

For $\alpha < 1$ the jump is still at the same position but \mathcal{N} increases only by a polynomial factor as is shown in Figure 7.1 for $\alpha = 0.25$.

Additionally this figure shows that the order of the largest component jumps from approximately the size of a single feature clique (which is concentrated around pn , see (7.6)) as a trivial lower bound to the order of the largest component to approximately the sum of the sizes of all feature cliques (which is for the same reasons concentrated around pmn) which is an upper bound to \mathcal{N} .

7.2 Branching Processes

In order to discover components in a graph we will use branching processes (for an overview of the topic of branching processes and for references to those used in proofs see Athreya and Vidyashankar [1999]) similar to the proofs in Chapter 5 of Janson et al. [2000]. We will explore the component by starting at a single vertex, generating its neighbours as descendants in a branching process and then the second neighbourhood as their descendants and so forth.

Let the random variable X with binomial distribution $\text{Bi}(n, p)$ denote the number of descendants (neighbours) of an arbitrary vertex. The Galton-Watson branching process on the variable X has the following properties (see Theorem 5.1 and Example 5.2 and 5.3 in Janson et al. [2000]).

1. If $np \xrightarrow{n \rightarrow \infty} c < 1$ the branching process on X dies out a.a.s.
2. If $np \xrightarrow{n \rightarrow \infty} c > 1$ the branching process dies out with probability $\rho(c)$ where $\rho(c)$ is the unique solution of

$$\rho = \exp(c(\rho - 1)) \quad (7.5)$$

in the interval $(0, 1)$.

Thus the main complication in the proof is to overcome the limitations of the branching process which deals with an essentially unbounded domain in contrast to the limited number of vertices in the graph.

The discovery of neighbours is (in contrast to the process used in the $G_{n,p}$ model) a two step process. First we let the vertex discover its features and then the features find the vertices they are assigned to. The features and the vertices used in each step will be ignored in the further process which will slightly downsize the universal feature set and the vertex set. As we will see later this deviation will not affect the ongoing process very much.

7.3 The Evolution for $\alpha > 1$

This section contains the proof of the first two statements of Theorem 7.1. After giving a sharp concentration result on the number of features a single vertex may have, we closely follow the branching process method used in Janson et al. [2000] to prove the results on the order of the largest component.

7.3.1 The Size of the Feature Set

In order to give precise estimates on the number vertices which get discovered by the branching process we need sharp bounds on the size of the feature set of a vertex. This result is similar to Lemma 6.2 but it needs to work even for graphs where we removed parts which are no longer available to the branching process.

Lemma 7.2. *Let v be a fixed vertex in a random intersection graph $G_{n,m,p}$ with $pn = o(1)$ and $p^2mn = \Theta(1)$. Furthermore let $W' \subseteq W$ be a subset of the universal feature set of size at least $m - 2pmn$ and $X_v := |W_v \cap W'|$ denote the random variable counting the number of features of v in W' . Then X_v is very likely close to its expectation or precisely:*

$$\mathbb{P} \left[|X_v - pm| > (pm)^{\frac{3}{4}} \right] \leq \exp \left(-\frac{(pm)^{\frac{1}{2}}}{3} \right)$$

Proof. For the expected number of features selected in W' we have $\mu := \mathbb{E}[X_v] \geq p(m - 2pmn) = pm - O(1)$ and $\mu \leq pm$.

Since the features are selected independently uniformly at random we can use Chernoff inequalities (6.7) and (6.8) to bound the deviation from the expected size.

$$\begin{aligned}
 \mathbb{P}\left[Y \geq pm + (pm)^{\frac{3}{4}}\right] &\leq \mathbb{P}\left[Y \geq \mu + (pm)^{\frac{3}{4}}\right] \\
 &\leq \exp\left(-\frac{(pm)^{\frac{3}{2}}}{2\left(\mu + (pm)^{\frac{3}{4}}/3\right)}\right) \\
 &\leq \exp\left(-\frac{(pm)^{\frac{3}{2}}}{2\left(pm + (pm)^{\frac{3}{4}}/3\right)}\right) \\
 &\leq \frac{1}{2} \exp\left(-\frac{(pm)^{\frac{1}{2}}}{3}\right)
 \end{aligned}$$

And for the lower tail using (6.8):

$$\begin{aligned}
 \mathbb{P}\left[Y \leq pm - (pm)^{\frac{3}{4}}\right] &= \mathbb{P}\left[Y \geq \mu + O(1) - (pm)^{\frac{3}{4}}\right] \\
 &\leq \exp\left(-\frac{\left((pm)^{\frac{3}{4}} - O(1)\right)^2}{2(pm - O(1))}\right) \\
 &\leq \frac{1}{2} \exp\left(-\frac{(pm)^{\frac{1}{2}}}{3}\right)
 \end{aligned}$$

Notice that these calculations (and thus the probability for the tails) remain valid even if we remove no features at all.

From the two tails above we may easily conclude the statement of the lemma. \square

7.3.2 Proof of Theorem 7.1, (7.1) and (7.2)

Proof of (7.1). We prove that for $c < 1$ the branching process starting at an arbitrary vertex v discovering all the vertices one by one will finish in at most $\frac{9 \ln n}{(1-c)^2}$ steps.

From Lemma 7.2 we know that there is with high probability no large deviation from the expected value in the size of a feature set. Our branching process starting at v now proceeds as follows. At first v discovers its features. If there are too many or too few of them (in the sense of Lemma 7.2) we abort.

Otherwise we let the features discover the vertices which hold them. Since the feature set of v has size $(1 + o(1))pm$ the probability for an individual vertex w to hold at least one feature in this set is

$$\mathbb{P}[\{v, w\} \in E(G_{n,m,p})] = 1 - (1 - p)^{(1+o(1))pm} \stackrel{(6.3)}{=} (1 + o(1))p^2m$$

and the neighbours of v will be chosen independently with this probability. Thus the expected number of new neighbours discovered will be:

$$\mathbb{E}[d(v)] \leq n(1 + o(1))p^2m$$

Now we remove W_v (the feature set of v) from the universal feature set and continue with discovering the features of the neighbours of v the same way we discovered the features of v and so on. We do this at most n times (only n vertices available) thus the probability that we will abort at any step because of the wrong size of the feature set is (due to Lemma 7.2) bounded by

$$n \exp\left(-\frac{(pm)^{\frac{1}{2}}}{3}\right) \xrightarrow{n \rightarrow \infty} 0.$$

Furthermore we did remove at most $n(1 + o(1))pm < 2pmn$ features from the universal feature set thus Lemma 7.2 was applicable all the time.

Observe that the probability that v is in a component of order at least k is bounded by the probability that the sum of the degrees of k vertices discovered in the process is at least $k - 1$. Since all features were discovered independent from earlier ones and thus all vertices were discovered in an independent manner, the probability for a component of order at least $k \geq \frac{9 \ln n}{(1-c)^2}$ can be bounded using a Chernoff inequality again. Let Y_i denote the number of neighbours of the i th vertex discovered in the process and notice that the expected value for the sum over the Y_i is bounded from above by $(1 + o(1))kp^2mn \leq kc'$ for $c' := \frac{c+1}{2}$.

$$\begin{aligned} n\mathbb{P}\left[\sum_{i=1}^k Y_i \geq k - 1\right] &= n\mathbb{P}\left[\sum_{i=1}^k Y_i \geq kc' + (1 - c')k - 1\right] \\ &\leq n \exp\left(-\frac{((1 - c')k - 1)^2}{2(c'k + (1 - c')k/3)}\right) \\ &\leq n \exp\left(-\frac{(1 - c')^2}{2}k\right). \end{aligned}$$

Resubstituting c' and k shows that this term tends to 0 as n tends to infinity which proves by (6.5) the theorem. \square

For the appearance of a giant component when $c > 1$ we will study the same branching process again using the proof of Janson et al. [2000].

Proof of (7.2). We start by proving that there is a.a.s. no component which has more than $k_- := \frac{50c}{(c-1)^2} \ln n$ or less than $k_+ := n^{2/3}$ vertices by proving the harder result that for every $k_- < k < k_+$ there are a.a.s. $\frac{(c-1)}{2}k$ vertices which are to be examined (have been discovered as neighbours but were not examined themselves). To prove this we have to look at no more than $k + \frac{c-1}{2}k = \frac{c+1}{2}k$ vertices.

Because of this we exclude in each step at most $\frac{c+1}{2}k_+$ vertices from the further process. Furthermore we do still downsize the universal feature set only for a very small amount for each vertex which discovers its neighbours as in the proof of (7.1). This gives independence for all steps of the branching process and thus one can bound the number of neighbours a vertex discovers from below by independent random variables

$Y_i^* \in \text{Bi}(n - \frac{c+1}{2}k_+, p'^2m)$ with p' such that $p'^2mn = \frac{3c+1}{4}$. The value for p' results from the lower bound on the size of feature set given by Lemma 7.2.

Now we can bound the probability of dying out after k steps or having too few discovered (but unexamined) vertices by the probability that

$$\sum_{i=1}^k Y_i^* \leq k - 1 + \frac{c-1}{2}k$$

Now the existence of such a process can be bounded by Chernoff inequality (6.8) and we get with $\mu := \mathbb{E} \left[\sum_{i=1}^k Y_i^* \right] = \frac{3c+1}{4}k - o(k)$ for $k_- \leq k \leq k_+$ and n large enough:

$$\begin{aligned} n \sum_{k=k_-}^{k_+} \mathbb{P} \left[\sum_{i=1}^k Y_i^* \leq k - 1 + \frac{c-1}{2}k \right] &= n \sum_{k=k_-}^{k_+} \mathbb{P} \left[\sum_{i=1}^k Y_i^* \leq \mu - \left(\frac{c-1}{4}k - o(k) + 1 \right) \right] \\ &\leq n \sum_{k=k_-}^{k_+} \exp \left(\frac{- \left(\frac{c-1}{4}k - o(k) + 1 \right)^2}{\frac{3c+1}{2}k} \right) \\ &\leq n \sum_{k=k_-}^{k_+} \exp \left(\frac{- \left(\frac{c-1}{4} \right)^2 k}{3c} \right) \\ &\leq nk_+ \exp \left(\frac{- \left(\frac{c-1}{4} \right)^2 k_-}{3c} \right) \end{aligned}$$

Because of the values for k_- and k_+ given at the beginning of the proof this tends to 0 as n tends to infinity and thus by (6.5) there is a.a.s.no process stopping between k_- and k_+ .

If there exist two different components T and U with $|T| \geq k_+$ and $|U| \geq k_+$ their sets of features W_T and W_U have to be disjoint. According to Lemma 7.2 a.a.s. $|W_U| \geq k_+ \frac{pm}{2}$. Thus the probability of disjointness is:

$$(1-p)^{k_+^2 \frac{pm}{2}} \stackrel{(6.4)}{\leq} \exp \left(-k_+^2 \frac{p^2m}{2} \right) = \exp \left(-n^{\frac{4}{3}} \frac{c}{2n} \right) \xrightarrow{n \rightarrow \infty} 0$$

Now we have that there is a.a.s.only one component with at least k_+ vertices, it remains to show that it has linear order. Let Y denote the number of vertices in components of order at most k_- . Let for each vertex $i \in V$ Y_i be the indicator variable for being in such a small component. We estimate the expectation and variance of Y .

For a single vertex the probability of being in a small component can be bounded from above and from below by the extinction probabilities of branching processes with distribution $\text{Bi}(n - k_-, (1 - o(1))p^2m)$ and $\text{Bi}(n, (1 + o(1))p^2m)$. The $o(1)$ terms in the two cases bound the possible deviations in the size of feature sets according to Lemma 7.2. By (7.5) we know that the probability of extinction of these two processes is ρ which results by linearity of expectation into $\mathbb{E}[Y] = \rho(c)n$.

In order to prove the concentration of Y around its expectation, we calculate its variance, or precisely using (6.6) we show that $\mathbb{E}[Y^2] = (1 + o(1))\mathbb{E}[Y]^2$. Two vertices being simultaneously in a small component is an event which occurs either if they are in the same component, in which case the probability can be bounded by the extinction probability for this component or they are in two components which means two extinctions have to occur independently.

$$\begin{aligned}\mathbb{E}[Y^2] &= \mathbb{E}\left[\left(\sum_{i=1}^n Y_i\right)^2\right] = \sum_{i,j} \mathbb{E}[Y_i Y_j] \\ &\leq n\rho(np)k_- + n\rho(np)n\rho((n - k_-)p) \\ &= (1 + o(1))n^2\rho(np)^2 = (1 + o(1))\mathbb{E}[Y]^2\end{aligned}$$

By Tschebyscheff's inequality (6.6) we can conclude that the number of small vertices is a.s. $\rho(c)n$ hence the largest component is of order $(1 - \rho(c))n$. \square

One further consequence of this proof is that for $\alpha > 1$ and $c > 1$ we can bound the order of the second largest component by $\frac{50c}{(c-1)^2} \ln n$.

7.4 The Evolution for $\alpha < 1$

If we have a small upper bound for the number of vertices two feature cliques have in common we can simply add the clique sizes (provided we know they are connected) in order to estimate the component order. This bound is the content of the following lemma.

Lemma 7.3. *Let Y be the random variable counting the number of vertices having more than one feature in a random intersection graph $G_{n,m,p}$ with $m := n^\alpha$ and $\alpha < 1$. Then for $p^2 m^2 n \gg \ln n$:*

$$\mathbb{P}[Y > 2p^2 m^2 n] \xrightarrow{n \rightarrow \infty} 0$$

and for $p^2 m^2 n \xrightarrow{n \rightarrow \infty} 0$:

$$\mathbb{P}[Y > 0] \xrightarrow{n \rightarrow \infty} 0$$

Proof. For a single fixed vertex v the probability of having more than one feature is (when $pm \rightarrow 0$):

$$\mathbb{P}[|W_v| > 1] = 1 - (1 - p)^m - (mp(1 - p))^{m-1} \stackrel{(6.3)}{=} (1 + o(1))m^2 p^2.$$

Since all vertices choose their features independently Y is a binomially distributed variable with expectation $nm^2 p^2$ and the second statement of the lemma follows by Markov inequality. For the first statement we can bound the deviation using Chernoff inequality (6.7).

$$\mathbb{P}[Y > 2p^2 m^2 n] \leq \mathbb{P}[Y > 2\mathbb{E}[Y]] \leq \exp\left(-\frac{3nm^2 p^2}{8}\right) \xrightarrow{n \rightarrow \infty} 0.$$

\square

Now we can start proving the component evolution for $\alpha < 1$.

Proof of (7.3). In order to reuse the results of Section 7.3 we interchange the role of the feature set and the vertex set and look at the largest component in the feature set instead of one in the vertex set. As we know from Theorem (7.1) there will be no component containing more than $\frac{9}{(1-c)^2} \ln m$ features. Exploiting again the symmetry between feature set and vertex set, we can use Lemma 7.2 to deduce that for every feature w

$$V_w = (1 + o(1))pn \quad (7.6)$$

with probability at least $1 - m \exp(-(pn)^{1/2}/3) = 1 - o(1)$. We can conclude that the order of the largest component is a.a.s. bounded by

$$\frac{9}{(1-c)^2} \ln m \cdot (1 + o(1))pn \leq \frac{10\sqrt{c}}{(1-c)^2} \sqrt{\frac{n}{m}} \ln m.$$

□

Proof of (7.4). We use the same method as in the last proof. With exactly the same argument we already have a.a.s. an upper bound for the order of the largest component of

$$(1 - \rho(c))m \cdot (1 + o(1))pn \leq (1 + o(1))\sqrt{c}(1 - \rho(c))\sqrt{mn}.$$

The lower bound can be achieved because the order of the component can be bound by the sum over the sizes of all cliques minus the number of vertices which occur in more than one clique multiplied with the multiplicity they occur. Or more precisely (with W_L denoting the set of features in the giant component in W and V_L denoting the vertices linked to it):

$$\begin{aligned} |V_L| &= \sum_{w \in W_L} |V_w| - \sum_{v \in V_L, |W_v| > 1} (|W_v| - 1) \\ &\geq (1 - \rho(c))m(1 + o(1))pn - \sum_{v \in V_L, |W_v| > 1} \max_{v \in V} \{|W_v|\} \end{aligned}$$

The probability of the existence of a vertex with more than $\ln m$ features is bounded by $n(pm)^{\ln m}$ which tends to 0 for our choice of p . Furthermore we know from Lemma 7.3 that there are at most $2p^2m^2n = 2cm$ vertices with more than one feature. Therefore

$$\begin{aligned} |V_L| &\geq (1 - \rho(c))m(1 + o(1))pn - 2cm \ln m \\ &= (1 + o(1))(1 - \rho(c))\sqrt{cmn} - 2cm \ln m \\ &= (1 + o(1))(1 - \rho(c))\sqrt{cmn}. \end{aligned}$$

□

As a direct consequence of this bound and the remark after the proof of (7.2) we have that for $\alpha < 1$ and $c > 1$ we can bound the order of the second largest component by

$$\frac{51c}{(c-1)^2} \ln m pn = \frac{51c\sqrt{c}}{(c-1)^2} \sqrt{\frac{n}{m}} \ln m.$$

7.4.1 Feature Cliques as Components

Similar to the evolution of $G_{n,p}$, which has lots of isolated vertices for very small p , there are stages of the evolution of $G_{n,m,p}$ where the feature cliques do not intersect. At this stage the component structure of $G_{n,m,p}$ is not very complex.

Proposition 7.4. *Let $G_{n,m,p}$ be a random intersection graph with $m := n^\alpha$ and $\alpha < \frac{1}{2}$ and $\ln n \ll pn \ll \frac{\sqrt{n}}{m}$. Then a.a.s. there are m components which are (feature) cliques and the rest of the graph consists of isolated vertices and thus a.a.s. $\mathcal{N}(G_{n,m,p}) = (1 + o(1))pn$.*

Proof. The statement follows directly from Lemma 7.3 and (7.6) because if there are no vertices with more than one feature there are only isolated vertices and feature cliques. \square

Chapter 8

Clique cover and feature reconstruction

8.1 Results

The main aim of this chapter is to develop and analyze simple algorithms which, given an intersection graph, quickly reproduce the underlying feature cliques. As the features of a network are likely to reflect important properties of the data, they represent important meta-information that will help in clustering, storing and searching it efficiently. An immediate example for such feature cliques are communities in the world wide web which share *common topics* and thus their webpages (represented by vertices) are highly interconnected via hyperlinks (represented by edges).

Since every graph can be seen as an intersection graph with the universal feature set being large enough, we want to (re)produce a universal feature set that is as small as possible. This is equivalent to the NP-hard problem of constructing an (edge) clique cover with a minimum number of cliques for the graph, see Garey and Johnson [1979], and hence we cannot expect to find an efficient algorithm which always finds an optimal solution. Instead, we present a simple greedy heuristic that constructs a generator of a given graph. Our main contribution is to prove that this algorithm performs *a.a.s.* optimally (this means with probability tending to one as n tends to infinity), when the input graph is chosen at random from our model $G_{n,m,p}$ for certain ranges of p . More precisely, we will prove the following two theorems.

Theorem 8.1. *Let a positive constant $\alpha < 1$, $n, m := n^\alpha$ and $\frac{\ln^2 n}{n} \leq p = O(\frac{1}{m})$ be given and let $G := G_{n,m,p} = (V, E)$ be a random intersection graph with $n = |V|$. Then there exists an algorithm which *a.a.s.* finds a bipartite graph $B = (V \cup W, A)$ with $|W| \leq m$ and $B^2[V] = G$ (a generator of G). Its running time is bounded by $O(n|E|)$.*

Theorem 8.2. *Let a positive constant $\alpha < 1$, $n, m := n^\alpha$ and $\frac{\ln^2 n}{n} \leq p < \min\{\frac{1}{5}m^{-\frac{2}{3}}, \frac{n}{8m^2}\}$ be given and let $G := G_{n,m,p} = (V, E)$ be a random intersection graph. Then there exists an algorithm which *a.a.s.* finds in polynomial time a bipartite graph $B = (V \cup W, A)$ with $|W| \leq m$ and $B^2[V] = G$ (a generator of G).*

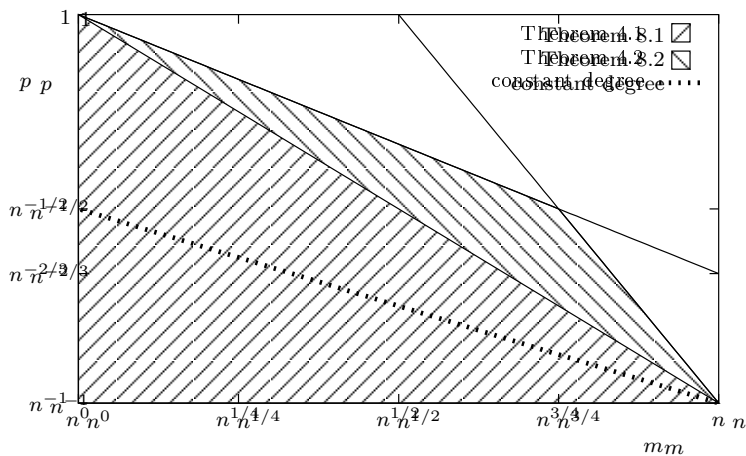


Figure 8.1: Ranges for p and m for which we prove the a.a.s.optimality of Algorithm 1

Notice that Theorem 8.2 covers a greater range of p at the expense of a larger (but still polynomial) running time of the algorithm. Observe that in particular graphs with constant expected degree (which seems appropriate for many real-world networks) are already covered by Theorem 8.1 and can thus be analyzed very efficiently. Figure 8.1 illustrates the range of m and p for which our theorems hold.

Following Guillaume and Latapy [2004], who compared real complex networks with random intersection graphs, we ran our algorithm on the same or similar real-world networks to obtain a clique cover. The results can be found in Chapter 10. The simulation results show that even very large graphs can be covered quite well with a reasonable number of cliques and a good running time.

This chapter is organized as follows. Section 8.2 contains the algorithm that gives rise to the theorems. In Section 8.3 we prove Theorem 8.1 which is just a warmup for the proof of Theorem 8.2 in Section 8.4.

8.2 The Algorithm

The following algorithm finds cliques in a graph by testing the common neighborhood of vertex subsets of fixed size k for completeness. From the cliques found in this way it takes the largest ones in order to cover the graph.

We shall use the following (slightly non-standard) notation: For the set $A \cup \{x\}$ we write $A+x$. Denote by $\Gamma(v)$ the set of vertices having edges to v and by $N(v) := \Gamma(v) + v$ the same set including v itself. For a vertex set U we denote by $Z(U)$ the common neighborhood of the vertices in U ($Z(U) := \bigcap_{i=1}^k N(v_i)$).

Algorithm 1.

Input: Graph $G = (V, E)$ on n vertices, $k \in \mathbb{N}$

Output: (partial) edge clique cover \mathcal{M} of G

FEATUREFIND(G, k)

- (1) $\mathcal{L} := \emptyset;$
- (2) **foreach** $U_k = \{v_1, \dots, v_k\} \subseteq V$
- (3) $Z = Z(U_k) := \bigcap_{i=1}^k N(v_i)$
- (4) **if** $G[Z]$ complete
- (5) $\mathcal{L} := \mathcal{L} + Z;$
- (6) $Y := \emptyset;$
- (7) **foreach** $Z \in \mathcal{L}$ in decreasing cardinality $|Z|$
- (8) **if** $E(G[Z]) \not\subseteq Y$
- (9) $Y := Y \cup E(G[Z]);$
- (10) $\mathcal{M} := \mathcal{M} + Z;$

We will use this algorithm with $k = 1$ to prove Theorem 8.1 and with larger k to prove Theorem 8.2. The set \mathcal{M} found by the algorithm contains the vertex sets seen by the individual features and can thus be considered as a subset of the feature set W of a possible generator of G .

The running time of the algorithm is clearly dominated by checking the clique property for the neighborhood of all k -subsets of V . Since the clique property can clearly be checked in time $O(|E|)$ this leads to a total of $O(\binom{n}{k}|E|)$. The following proposition gives rise to an algorithm which needs much less time in practice.

Proposition 8.3. *Let $G = (V, E)$ be a graph and let $U \subseteq V$ be such that $C := Z(U) = \bigcap_{u \in U} N(u)$ is a clique in G . Furthermore let U' be an arbitrary subset of C . If $Z(U')$ is a clique then $Z(U') = C$.*

Proof. Since C is a clique it is immediate that for every subset $U' \subseteq C$ all vertices of C are adjacent to all vertices of U' , hence $C \subseteq Z(U')$. Now assume that $Z(U')$ is a clique and that there is a vertex v in $Z(U')$ which is not in C . Since $C \subseteq Z(U')$ all vertices in C (and especially in U) are adjacent to v but this means $v \in Z(U) = C$ which contradicts the assumption that $v \notin C$. Thus v cannot exist and the statement is proven. \square

This proposition implies that every set U_k which is a subset of a clique that has been found in an earlier stage of the algorithm does not have to be checked anymore, which in practice reduces the number of sets to be checked dramatically.

Furthermore note that for $k = 1$ (and in fact even for $k = 2$) sorting the cliques (starting at line 7) and taking only the largest ones is not necessary which speeds up the algorithm as well.

Proposition 8.4. *Let $G = (V, E)$ be a graph and let $e = (u, v) \in E$ be such that $C := N(u) \cup N(v)$ is a clique in G or let v be such that $C := N(v)$ is a clique in G . Then every minimum edge clique cover of G contains a subset of C .*

Proof. This is a simple corollary of Proposition 8.3 since there is no clique in G which contains e (resp. v) and is not a subset of C . \square

That means for $k = 1$ and $k = 2$ the size of the computed clique cover is a lower bound to the clique cover number of the graph. In order to improve further on this lower bound in the experiments we decided to let the algorithm work iteratively, details can be found in Section 10.3.

8.3 The case $k = 1$

We first show that almost surely every feature clique contains a vertex with only one feature.

Lemma 8.5. *Let $G_{n,m,p}$ with $m := n^\alpha$, $\alpha < 1$ and $\frac{\ln^2 n}{n} \leq p = O(\frac{1}{m})$ be a random intersection graph. Then a.a.s. every feature clique V_w contains a vertex for which w is the only feature:*

$$\forall w \in W \exists v \in V_w : W_v = \{w\}.$$

Proof. For $p \geq \frac{\ln^2 n}{n}$ we know from Lemma 6.1 that we can condition on the event that there is a.a.s. no feature clique with less than $\frac{pn}{2}$ vertices. Now fix a single feature w , let it choose its clique V_w and determine the probability that all the vertices inside V_w choose another feature. Summing over all features we can then bound the probability for the existence of such a w by

$$\begin{aligned} \mathbb{P}[\exists w \in W, \forall v \in V_w : |W_v| > 1] &\leq m \left(1 - (1 - p)^{m-1}\right)^{\frac{pn}{2}} \\ &\stackrel{(6.4)}{\leq} m \left(1 - e^{-2pm}\right)^{\frac{pn}{2}} \\ &\leq m \left(1 - e^{-O(1)}\right)^{\frac{pn}{2}} \\ &\leq m \left(1 - e^{-O(1)}\right)^{\frac{\ln^2 n}{2}}. \end{aligned}$$

This tends to 0 because for n large enough $\ln(1 - e^{-O(1)}) \ln n < -\alpha$. □

Theorem 8.1 now follows immediately from this lemma because Algorithm 1 only needs to “find” the vertex from Lemma 8.5 which it will surely achieve running with $k = 1$.

Proof of Theorem 8.1. We run Algorithm 1 with $k = 1$ and claim that a.a.s. the produced list \mathcal{L} will contain exactly the feature cliques. By Lemma 8.5 we can assume that every feature clique V_w contains a vertex u_w for which w is the only feature ($W_{u_w} = \{w\}$). Observe that for such a vertex u_w the neighborhood $N(u_w)$ is a feature clique. This already implies that all feature cliques will be contained in \mathcal{L} .

Now assume that there is a vertex v with more than one feature (e.g. $x, y \in W_v$). Since u_x and u_y must lie in $N(v)$ (because v shares one feature with each of them) and since there is no edge between u_x and u_y (they have only one feature) $N(v)$ cannot be a clique. Thus if $N(v)$ is a clique, then this implies that $v = u_w$ for some feature w , and therefore \mathcal{L} contains exactly the feature cliques.

The running time is bounded from above by the time needed to check the clique property for at most n sets which can surely be done in $O(n|E|)$. \square

Theorem 8.1 already covers a significant portion of interesting intersection graphs, in particular graphs with expected constant degrees (linear number of edges) and with a giant component. Both properties occur when $p = c/\sqrt{mn}$ (see Chapter 7 for details).

8.4 The case $k > 1$

The proof of Theorem 8.2 needs some more lemmas because the a.a.s.existence of a vertex with only one feature cannot be guaranteed for larger p . We will use two other asymptotic properties of the feature cliques instead. First we prove that feature cliques are maximal with respect to inclusion (Lemma 8.6) and from this deduce that in fact there are no larger cliques in the graph (Lemma 8.8). Together with the a.a.s.existence of at least one set U_k whose common neighborhood $Z(U_k)$ is complete (Lemma 8.7) this will prove the theorem.

Lemma 8.6. *Consider $m := n^\alpha$, $\alpha < 1$, a positive constant k and a random intersection graph $G_{n,m,p}$ with $\frac{k}{m} \leq p < \frac{1}{\sqrt{m \ln n}}$. Then a.a.s.every feature clique is inclusion maximal:*

$$\forall w \in W \forall v \in V : V_w \not\subseteq \Gamma(v).$$

Proof. First observe that the statement of the formula is trivial for $v \in V_w$ since no vertex can be part of its own neighborhood. Now assume that we have the bounds on the sizes of the feature cliques and sets from Lemma 6.1 and Lemma 6.2. Suppose that for some vertex w there exists a vertex $v \notin V_w$ with $V_w \subseteq \Gamma(v)$. We will show that the probability of this event vanishes. First consider the case where $pm > 3 \ln n$:

$$\begin{aligned} \mathbb{P}[\exists w \in W, v \in V : V_w \subseteq \Gamma(v)] &\leq mn \sum_{i=1}^{2pm} \binom{m}{i} p^i (1 - (1-p)^i)^{\frac{pn}{2}} \\ &\stackrel{(6.1)(6.3)}{\leq} mn \sum_{i=1}^{2pm} \left(\frac{emp}{i}\right)^i (pi)^{\frac{pn}{2}} \\ &\leq mn \sum_{i=1}^{2pm} \left(\frac{emp}{i}\right)^{\frac{pn}{2}} (pi)^{\frac{pn}{2}} \quad \text{with } i < emp < \frac{pn}{2} \\ &\leq mn 2pm (emp^2)^{\frac{pn}{2}} \\ &\leq mn 2pm \left(\frac{e}{\ln^2 n}\right)^{\frac{pn}{2}}, \end{aligned}$$

which tends to 0 because $\frac{e}{\ln^2 n} \rightarrow 0$ and $pn \geq n^{1-\alpha}$.

Now for the case where $pm \leq 3 \ln n$:

$$\begin{aligned}
 \mathbb{P}[\exists w \in W, v \in V : V_w \subseteq \Gamma(v)] &\leq mn \sum_{i=1}^{21 \ln n} \binom{m}{i} p^i (1 - (1-p)^i)^{\frac{pn}{2}} \\
 &\stackrel{(6.2)(6.3)}{\leq} mn \sum_{i=1}^{21 \ln n} (mp)^i (pi)^{\frac{pn}{2}} \\
 &\leq mn \sum_{i=1}^{21 \ln n} (p^2 mi)^{\frac{pn}{2}} \\
 &\leq 21mn \ln n (21p^2 m \ln n)^{\frac{pn}{2}} \\
 &\leq 21mn \ln n \left(\frac{21}{\ln n} \right)^{\frac{pn}{2}},
 \end{aligned}$$

which tends to 0 because $\frac{21}{\ln n} \rightarrow 0$ and $pn \geq n^{1-\alpha}$. \square

Now we prove that we can indeed find the feature cliques with our algorithm.

Lemma 8.7. *Let $\varepsilon > 0$ be fixed and consider $m := n^\alpha$, $\alpha < 1$, an integer $k > \frac{\alpha+1}{2\alpha\varepsilon}$ and a random intersection graph $G_{n,m,p}$ with $\frac{k}{m} \leq p < m^{-\frac{1}{2}-\varepsilon}$. Then a.a.s. every feature clique has a subset U_k of size k such that $V_w = Z(U_k)$ (with Z being defined in the algorithm).*

Proof. Fix a feature w and let U_k be a fixed k -clique with $U_k \subseteq V_w$ (remember that all subsets of V_w are cliques). Furthermore let $v \in V_w$ be an arbitrary vertex. As V_w is a clique, $U_k \subseteq N(v)$ which is equivalent to $v \in \bigcap_{i=1}^k N(u_i) = Z(U_k)$. Thus $v \in V_w$ and, because v was chosen arbitrarily, $V_w \subseteq Z(U_k)$. If $Z(U_k)$ is complete we know from Lemma 8.6 that $Z(U_k) = V_w$ and we are done.

So assume the opposite, e.g. there are $x, y \in Z(U_k)$ which are not adjacent. Since V_w is a clique, x or y has to be outside of V_w . Let us assume it is x , then the event of $Z(U_k)$ being not complete implies the event that there exists an $x \in Z(U_k) \setminus V_w$. This means there is an x that is in the neighborhood of all vertices in U_k but does not see feature w .

We bound the probability for this event by summing over all possible sets of (at most k) features which connect x and U_k .

$$\begin{aligned}
 \mathbb{P}[\exists x \in V \setminus V_w \forall u \in U_k : x \in \Gamma(u)] &\leq n \sum_{i=1}^k \binom{m}{i} p^i (1 - (1-p)^i)^k \\
 &\stackrel{(6.1)(6.3)}{\leq} n \sum_{i=1}^k \left(\frac{epm}{i} \right)^i (pi)^k \\
 &\leq n \sum_{i=1}^k (ep^2 m)^k \quad \text{with } i \leq k \leq pm \\
 &= nk(ep^2 m)^k.
 \end{aligned}$$

If this tends to 0, a subset U_k will a.a.s. have $Z(U_k) = V_w$ for our fixed w . In order to have this for all w , we check that

$$mnk(ep^2m)^k < mnk(em^{-2\varepsilon})^k = ke^k n^{\alpha+1-2\varepsilon k\alpha} \rightarrow 0,$$

which happens indeed for $k > \frac{\alpha+1}{2\alpha\varepsilon}$. \square

Finally we state that the sorting step at the end of the algorithm will indeed list the feature cliques first. In order to do so, we prove that a.a.s. all large cliques are feature cliques.

Lemma 8.8. *Consider a random intersection graph $G_{n,m,p}$ with $m := n^\alpha$ and $\frac{k}{m} \leq p < \min\{\frac{1}{5}m^{-\frac{2}{3}}, \frac{n}{8m^2}\}$ for some constant k . Then a.a.s. every clique of size at least $\frac{pn}{2}$ is a feature clique:*

$$\forall S \subseteq V \text{ with } |S| > \frac{pn}{2} \text{ and } G[S] \text{ is complete} : \exists w \in W \text{ such that } S \subseteq V_w.$$

Proof. Assume that the statement of the lemma is wrong. Thus there exists a clique S of size $\frac{pn}{2} + 1$ which is not a feature clique. Let $s \in S$ be an arbitrary vertex in S . Again, we first consider the case where $pm > 3 \ln n$. From Lemma 6.2 we know that a.a.s. no vertex in V has more than $2pm$ features, so this applies to s , too. But since s has $\frac{pn}{2}$ neighbors, there has to exist a subset $X \subseteq N(s)$ of size $\frac{pn}{4pm} = \frac{n}{4m}$ which shares a common feature w (by the pigeon hole principle). Furthermore there has to exist a vertex $v \in S$ with $v \notin V_w$, otherwise S would be inside a feature clique. We now bound the probability of the existence of such an X and v with $X \subseteq \Gamma(v)$ (remember that S is a clique). Here we use that by Lemma 6.1 the size of V_w is a.a.s. at most $2pn$ and by Lemma 6.2 $|W_v| \leq 2pm$.

$$\begin{aligned} & \mathbb{P} \left[\exists w \in W, v \in V, X \subseteq V_w : |X| = \frac{n}{4m} \wedge X \subseteq \Gamma(v) \right] \\ & \leq mn \binom{2pn}{|X|} \sum_{i=1}^{2pm} \binom{m}{i} p^i (1 - (1-p)^i)^{|X|} \\ & \stackrel{(6.1)}{\leq} mn \binom{2pn}{\frac{n}{4m}} \sum_{i=1}^{2pm} \left(\frac{emp}{i} \right)^i (1 - (1-p)^i)^{\frac{n}{4m}} \\ & \stackrel{(6.1)(6.3)}{\leq} mn(8epm)^{\frac{n}{4m}} \sum_{i=1}^{2pm} \left(\frac{emp}{i} \right)^i (pi)^{\frac{n}{4m}} \\ & \leq mn(8epm)^{\frac{n}{4m}} \sum_{i=1}^{2pm} \left(\frac{emp}{i} \right)^{\frac{n}{4m}} (pi)^{\frac{n}{4m}} \quad \text{with } i < 2pm < \frac{n}{4m} \\ & = mn(8epm)^{\frac{n}{4m}} 2pm(ep^2m)^{\frac{n}{4m}} \\ & = 2pm^2n(8e^2p^3m^2)^{\frac{n}{4m}} \\ & \leq 2pm^2n \left(\frac{72}{125} \right)^{\frac{n}{4m}}, \end{aligned}$$

which tends to 0.

For the case where $pm \leq 3 \ln n$ Lemma 6.2 only gives a bound of $21 \ln n$ on the size of the feature set. With the same considerations as above this leads to a set X of size $\frac{pn}{42 \ln n}$ and hence:

$$\begin{aligned}
 & \mathbb{P} \left[\exists w \in W, v \in V, X \subseteq V_w : |X| = \frac{pn}{42 \ln n} \wedge X \subseteq \Gamma(v) \right] \\
 & \leq mn \left(\frac{2pn}{42 \ln n} \right) \sum_{i=1}^{21 \ln n} \left(\frac{emp}{i} \right)^i \left(1 - (1-p)^i \right)^{\frac{pn}{42 \ln n}} \\
 & \stackrel{(6.1)(6.3)}{\leq} mn(84e \ln n)^{\frac{pn}{42 \ln n}} \sum_{i=1}^{21 \ln n} (mp)^i (pi)^{\frac{pn}{42 \ln n}} \\
 & \leq mn(84e \ln n)^{\frac{pn}{42 \ln n}} \sum_{i=1}^{21 \ln n} (p^2 mi)^{\frac{pn}{42 \ln n}} \\
 & \leq mn(84e \ln n)^{\frac{pn}{42 \ln n}} 21 \ln n (21p^2 m \ln n)^{\frac{pn}{42 \ln n}} \\
 & = 21mn \ln n (1764ep^2 m \ln^2 n)^{\frac{pn}{42 \ln n}} \\
 & \leq 21mn \ln n (80em^{-1/3} \ln^2 n)^{\frac{pn}{42 \ln n}},
 \end{aligned}$$

which tends to 0. □

The proof of Theorem 8.2 now merely requires collecting the statements of the lemmas.

Proof of Theorem 8.2. We make a case distinction over p . For $\frac{\ln^2 n}{n} \leq p = O(\frac{1}{m})$ we already know from Theorem 8.1 that the statement is true.

Now let $k := 6/\alpha$ and consider $\frac{k}{m} < p < \frac{1}{5}m^{-\frac{2}{3}}$. Set $\varepsilon = 1/6$ and apply Lemma 8.7: a.a.s. for each feature $w \in W$ there exists a set $U_k(w)$ with $Z(U_k(w)) = V_w$. Hence all feature cliques will be listed in \mathcal{L} after running the algorithm with k chosen as above.

Since we know from Lemma 6.1 that there is a.a.s. no feature clique with less than $\frac{pn}{2}$ vertices and from Lemma 8.8 that all cliques with more than $\frac{pn}{2}$ vertices are feature cliques we can conclude that sorting the list of cliques by their size and taking the elements until the graph is covered will a.a.s. succeed in reconstructing a bipartite graph which generates our input graph as an intersection graph.

Again the running time of our algorithm is bounded by the time needed to check the clique property for the joint neighborhood of all subsets of size k , and thus $O\left(\binom{n}{k}|E|\right)$. □

Chapter 9

Colouring heuristics and the clique number

9.1 Results

The aim of this chapter is to investigate the evolution of the chromatic number of $G_{n,m,p}$. As usual, denote by $\chi(G)$ the chromatic number of G and by $\omega(G)$ the size of the largest clique in G . The computation of these two fundamental parameters has long been known to be NP-hard. Our main results are that for a random intersection graph $G = G_{n,m,p}$ where m and p lie in a certain range, asymptotically almost surely $\chi(G)$ and $\omega(G)$ can be computed efficiently by simple colouring heuristics and actually coincide.

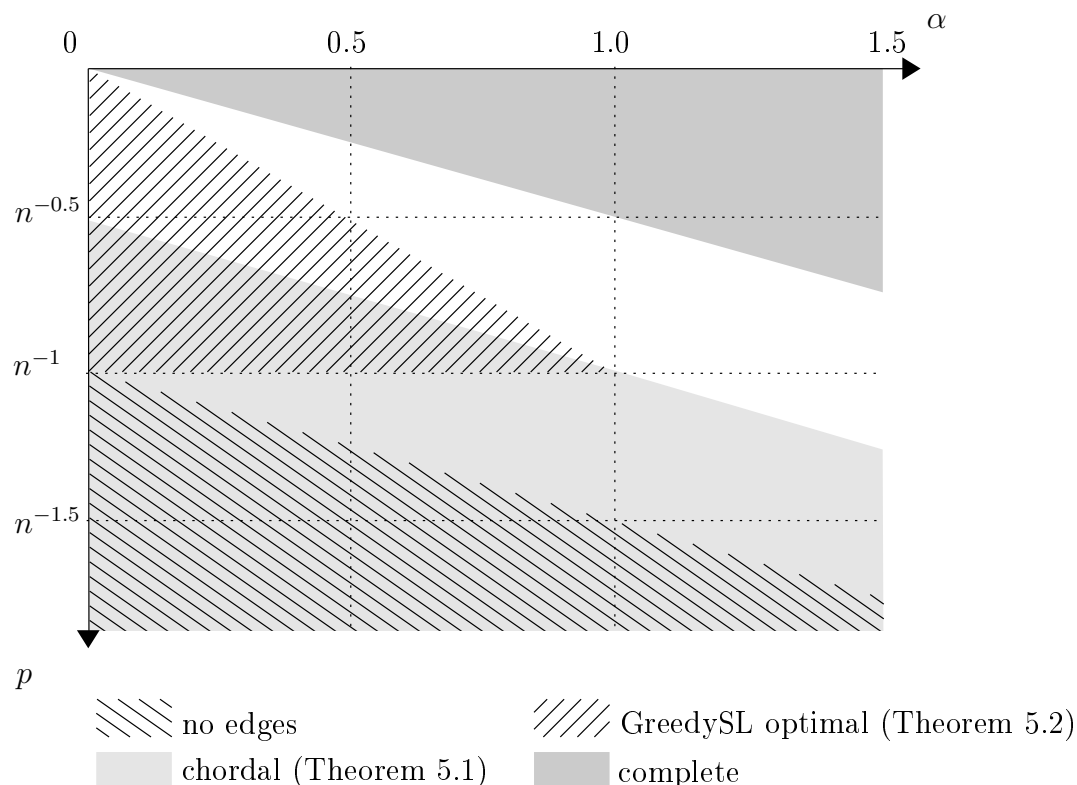
Theorem 9.1. *Let $m := n^\alpha$ with $\alpha > 0$ fixed and $p \ll \sqrt{\frac{1}{nm}}$. Then $G_{n,m,p}$ can a.a.s. be coloured optimally in linear time and $\chi(G_{n,m,p}) = \omega(G_{n,m,p})$.*

Theorem 9.2. *Let $m := n^\alpha$ with $0 < \alpha < 1$ fixed and $p \ll \frac{1}{m \ln n}$. Then $G_{n,m,p}$ can a.a.s. be coloured optimally in linear time. Moreover, for $np > \ln^4 n$ we have a.a.s.*

$$\chi(G_{n,m,p}) = \omega(G_{n,m,p}) \sim np.$$

Note that in principle one could also state in Theorem 9.1 that for $np > \ln^4 n$ we have a.a.s. $\chi(G_{n,m,p}) = \omega(G_{n,m,p}) \sim np$, but this is redundant since $np > \ln^4 n$ and $p \ll \sqrt{\frac{1}{nm}}$ together imply $\alpha < 1$ and thus the two theorems overlap in this case. Figure 9.1 gives an overview about the parameter ranges where our theorems apply together with some basic properties of random intersection graphs.

Applications. We have tested our colouring heuristics on real-world networks from application areas such as the internet, cooperation graphs and protein databases. Although we cannot prove that those networks can be modelled well with random intersection graphs having parameters in the range covered by our theorems, the heuristics described could colour those graphs optimally in many cases – see Section 10.4 for details. Still the question remains, *why* one should try to *colour* complex networks. Of course, knowledge of the chromatic number gives important structural information of a general nature, but

Figure 9.1: Ranges for p and α where we colour optimally

while for instance the clique number is practically meaningful – the size of the largest cluster in the network – the chromatic number seems to be of less immediate use.¹

There is however one important application of the chromatic number, and this is exactly the clique number. Suppose we have a heuristic that tries to find the maximal size of a clique. If we also have a heuristic that tries to determine the minimum number of colours, and both of the proposed numbers coincide (or are at least very close to each other), then this proves that both numbers have already reached (near-) optimal values. This is precisely what we did in our experiments: we applied different heuristics discussed in Chapter 8 to find large cliques (and good clique covers) in the networks. At the same time, we tried to find good colourings of real-world networks using the greedy algorithms discussed here. The results showed that, just as predicted for intersection graphs by Theorems 9.1 and 9.2, the proposed chromatic number and clique number indeed coincide in many cases.

¹One possible application, not to be taken too seriously, could be to distribute film-stars to a minimum number of hotels (colour classes) in such a way that co-stars of the same movie are not put in the same hotel, just to avoid trouble.

In a way this is very reminiscent of the theory of perfect graphs. In fact, $G_{n,m,p}$ with m and p as in Theorem 9.1 is a.a.s. perfect, and we can thus use some of the perfect graph methodology to give a short proof of the theorem. For parameters m and p as in Theorem 9.2, although $\chi(G_{n,m,p}) = \omega(G_{n,m,p})$ a.a.s., $G_{n,m,p}$ is not perfect and hence a different colouring strategy has to be used for this case.

9.2 Proofs

In the following two subsections we describe two simple and well known deterministic algorithms that find a proper colouring of a given input graph $G = (V, E)$ in linear time. Both algorithms are greedy heuristics: they colour the vertices in a prescribed order and assign to each vertex the smallest colour that has not been used for any of its neighbours which are already coloured. Thus the main task is to prove the following: if the input graph G is a random intersection graph $G_{n,m,p}$ with parameters n , m and p as given in Theorems 9.1 and 9.2, then these algorithms will asymptotically almost surely produce a colouring with (at most) $\omega(G)$ different colours. Hence the colouring is optimal and $\chi(G) = \omega(G)$, as required.

The additional claim in Theorem 9.2 that a.a.s. $\omega(G)$ is of order np will follow from the fact that the largest clique is a feature clique, which according to Lemma 6.1 is of that order.

9.2.1 Perfect Elimination Scheme

The aim of this subsection is to prove Theorem 9.1. Here is the basic idea of our colouring algorithm. We first try to order the vertices of the graph as x_n, \dots, x_1 in such a way that for every vertex x_i the ‘remaining neighbourhood’ $\Gamma(x_i) \cap \{x_{i-1}, \dots, x_1\}$ induces a clique in G . Having established this ordering, we greedily colour the vertices in the (reverse) order x_1, \dots, x_n . Observe that this implies that vertices which are contained in many different cliques, e.g. those that have many features, will be coloured relatively early.

Such an ordering is called a *perfect elimination scheme*, in short *PES*. Tarjan and Yannakakis [1984] proved that, if a graph has a PES, a so-called maximum cardinality search will produce a PES in linear time. If the graph doesn’t have a PES, then the procedure returns an arbitrary ordering. This leads to the following greedy colouring heuristic:

Algorithm 2.

Input: Graph $G = (V, E)$ on n vertices

Output: colouring of G

GREEDYCOLOURPES(G)

- (1) $A := \emptyset$
- (2) **for** $i := 1$ **to** n
- (3) choose $x_i \in V \setminus A$ such that $|\Gamma(x_i) \cap A|$ is maximal
- (4) $A := A + x_i$
- (5) **for** $i := 1$ **to** n
- (6) colour x_i with the smallest colour not occurring in $\Gamma(x_i)$

The following three crucial facts have been known for a long time:

1. a graph G has a PES (and it can be found in linear time and can be found as described above) if and only if G is *chordal*, i.e. it does not contain an induced cycle with more than three vertices Tarjan and Yannakakis [1984],
2. chordal graphs are perfect [Diestel, 1997, Chapter 5.5], thus in particular $\chi(G) = \omega(G)$, and
3. if a PES exists for G , then using it as described above the greedy colouring procedure colours G optimally.

The last observation is a folklore result and obviously true: if the set of the already coloured neighbours of every vertex x_i forms a clique when x_i is coloured, then whenever a vertex x_i needs a new colour k , we have just found a clique of size k , and hence k colours are really needed to colour the graph.

Now all that remains to do is to prove that $G_{n,m,p}$ is chordal for the given parameters n , m and p , which will be done in the following lemma.

Lemma 9.3. *Let $m := n^\alpha$ for $\alpha > 0$ fixed and $p \ll \sqrt{\frac{1}{nm}}$. Then $G_{n,m,p}$ is a.a.s. chordal.*

Proof. Let $G = G_{n,m,p}$ be a random intersection graph and $B = (V \cup W, E_B)$ a bipartite generator of G . By definition, G is chordal iff it does not contain an induced cycle of length at least four. Suppose that v_1, \dots, v_k form an induced cycle C_k in G . Then there must exist features w_1, \dots, w_k such that w_i is a feature of both v_i and v_{i+1} for all $i \in [k-1]$, and w_k is a feature for both v_k and v_1 . Moreover all the w_i are distinct, since otherwise the cycle wouldn't be induced. This yields a cycle $v_1, w_1, v_2, w_2, \dots, v_k, w_k$ in the generator B . The probability for such a cycle in B can obviously be bounded from above by p^{2k} , and multiplying this with the number of possibilities to choose v_1, \dots, v_k and w_1, \dots, w_k we get:

$$\mathbb{P}[G \text{ contains an induced } C_k] \leq n^k m^k p^{2k} = (nmp^2)^k.$$

The probability of G being not chordal is now bounded by:

$$\begin{aligned}
\mathbb{P}[G \text{ is not chordal}] &\leq \sum_{k=4}^{\min(n,m)} \mathbb{P}[G \text{ contains an induced } C_k] \\
&\leq \sum_{k=4}^{\min(n,m)} (nmp^2)^k \\
&\leq \sum_{k=0}^{\infty} (nmp^2)^k - 1 = \frac{1}{1 - nmp^2} - 1,
\end{aligned}$$

which tends to 0 for n tending to infinity because nmp^2 tends to 0. \square

A second moment calculation (see Singer [1995]) shows that $p = \sqrt{\frac{1}{nm}}$ is in fact the threshold function for the appearance of induced cycles of *fixed* length $k \geq 4$ in random intersection graphs. Thus for $p \gg \sqrt{\frac{1}{nm}}$ these graphs are a.a.s. not chordal.

9.2.2 Smallest Last Heuristic

The aim of this subsection is to prove Theorem 9.2. Again we employ a greedy strategy but this time the precomputed ordering x_1, \dots, x_n of the vertices is slightly different. Suppose we have already selected x_n, \dots, x_{i+1} . Then among the remaining vertices x_i is the vertex with the smallest number of neighbours (among the remaining vertices). More precisely:

Algorithm 3.

Input: Graph $G = (V, E)$ on n vertices

Output: colouring of G

GREEDYCOLOURSMALLESTLAST(G)

- (1) $A := V$
- (2) **for** $i := n$ **downto** 1
- (3) choose $x_i \in A$ such that $|\Gamma(x_i) \cap A|$ is minimal
- (4) $A := A - x_i$
- (5) **for** $i := 1$ **to** n
- (6) colour x_i with the smallest colour not occurring in $\Gamma(x_i)$

As there may be more than one such ordering, we denote by $\chi_{\text{SL}}(G)$ the maximum number of colours that GreedyColourSmallestLast(G) uses for an input graph G . It is well known [Diestel, 1997, Chapter 5.2] that the number of colours used by the algorithm is always bounded from above by the maximal minimum degree of all subgraphs of G , plus one:

$$\chi_{\text{SL}}(G) \leq 1 + \max_{H \subseteq G} \delta(H). \quad (9.1)$$

From this we derive the following simple proposition.

Proposition 9.4. *If G is a graph such that*

$$\text{every vertex } v \text{ has less than } \omega(G) \text{ neighbours of degree at least } \omega(G), \quad (9.2)$$

then

$$\chi_{SL}(G) = \omega(G) = \chi(G).$$

Proof. We claim that (9.2) implies that

$$1 + \max_{H \subseteq G} \delta(H) \leq \omega(G). \quad (9.3)$$

Suppose for a contradiction that there exists a subgraph H with $1 + \delta(H) > \omega(G)$. Let v be a vertex of minimal degree in H , i.e. $d_H(v) = \delta(H) \geq \omega(G)$. Then for *all* neighbours w of v in H we have

$$d_G(w) \geq d_H(w) \geq d_H(v) = \delta(H) \geq \omega(G),$$

and since there are $d_G(v) \geq d_H(v) = \delta(H) \geq \omega(G)$ neighbours of v in G , this contradicts the property in (9.2), which proves the claim in (9.3).

Now we are done, since

$$\chi(G) \leq \chi_{SL}(G) \stackrel{(9.1)}{\leq} 1 + \max_{H \subseteq G} \delta(H) \stackrel{(9.3)}{\leq} \omega(G) \leq \chi(G).$$

□

Let us move back to intersection graphs. In the following we call a vertex v *rich* if it has at least two features. Obviously, the only way that a vertex can have degree at least $\omega(G)$ is if it is rich. Hence we have the following corollary.

Corollary 9.5. *Suppose that G is an intersection graph such that every vertex has less than $\omega(G)$ rich neighbours, then*

$$\chi_{SL}(G) = \omega(G) = \chi(G).$$

□

In order to prove that in our random intersection graph, the condition of the above corollary is a.a.s. satisfied, we first obtain an upper bound on the number of rich vertices in each feature clique.

Lemma 9.6. *Let $m = n^\alpha$ for $0 < \alpha < 1$ fixed, $p \geq \frac{10 \ln^2 n}{n}$ and $t \geq 0$. Denote by ω_f the size of a largest feature clique in $G_{n,m,p}$. Then in a random intersection graph $G_{n,m,p}$ the probability that there exists a feature clique C with more than $\omega_f m p + t$ rich vertices is at most*

$$m \exp \left(- \frac{t^2}{2\omega_f m p + 2t/3} \right)$$

Proof. Let $C \subseteq V$ denote an arbitrary feature clique in G . For $v \in C$ we denote by $X_{C,v}$ the random variable which is 1 whenever v is rich and 0 otherwise. Then

$$\mathbb{P}[X_{C,v} = 1] = 1 - (1 - p)^{m-1} \stackrel{(6.3)}{\leq} 1 - (1 - (m-1)p) \leq mp.$$

Let $X_C := \sum_{v \in C} X_{C,v}$ count the rich vertices in C . For the expectation of X_C we have:

$$\mathbb{E}[X_C] = \sum_{v \in C} \mathbb{P}[X_{C,v} = 1] \leq \omega_f mp.$$

Using the Chernoff bound we get:

$$\begin{aligned} \mathbb{P}[X_C \geq \omega_f mp + t] &\leq \mathbb{P}[X_C \geq \mathbb{E}[X_C] + t] \\ &\stackrel{(6.7)}{\leq} \exp\left(-\frac{t^2}{2\mathbb{E}[X_C] + 2t/3}\right) \leq \exp\left(-\frac{t^2}{2\omega_f mp + 2t/3}\right). \end{aligned}$$

Of course the events ' $X_C \geq \omega_f mp + t$ ' are not independent of each other for overlapping feature cliques C , but using linearity of expectation and the Markov inequality (6.5) we can bound the probability of existence of a feature clique with too many rich vertices by the expression in the lemma. \square

Proof of Theorem 9.2. We want to apply Corollary 9.5 and hence need to show that in $G = G_{n,m,p}$ every vertex has less than $\omega(G)$ rich neighbours. Recall that $m := n^\alpha$ with $0 < \alpha < 1$ fixed and $p \ll \frac{1}{m \ln n}$. First observe that we can assume that $pn > \ln^4 n$, since otherwise p would be so small that we could apply Theorem 9.1 instead. Set

$$t := \max(3 \ln n, \sqrt{nmp^2 \ln n}),$$

and consider an arbitrary small $\varepsilon > 0$. We shall make use of the following two technical observations (involving t) that will be verified later:

$$21 \ln n((1 + \varepsilon)nmp^2 + t) \leq (1 - \varepsilon)np, \tag{9.4}$$

$$m \exp\left(-\frac{t^2}{2(1 + \varepsilon)nmp^2 + 2t/3}\right) \leq n^{\alpha-1}. \tag{9.5}$$

Again denote by ω_f the size of a largest feature clique in $G = G_{n,m,p}$ and consider the following events that have already been discussed in Lemmas 6.1, 6.2 and 9.6 respectively:

- \mathcal{A} : for all $w \in W : ||V_w| - pn| < \varepsilon pn$,
- \mathcal{B} : for all $v \in V : |W_v| \leq 21 \ln n$,
- \mathcal{C} : every feature clique C has at most $\omega_f mp + t$ rich vertices.

Let Y_v be the number of rich neighbours of a vertex v . Then Y_v is bounded from above by the number of feature cliques containing v , multiplied with the number of rich vertices per feature clique, and we can then compare this to the size of a feature clique, which is a lower bound for $\omega(G)$. So if all the events $\mathcal{A}, \mathcal{B}, \mathcal{C}$ hold, then

$$Y_v \leq 21 \ln n ((1 + \varepsilon)pn \mp t) \stackrel{(9.4)}{\leq} (1 - \varepsilon)np \stackrel{(\mathcal{A})}{<} \omega_f - 1 < \omega(G), \quad (9.6)$$

which would immediately prove (most of) the statements in Theorem 9.2 because of Corollary 9.5. To prove that $\omega(G) \sim np$, note that by the estimate in (9.6) there is no vertex v with $\omega_f - 1$ rich neighbours, and hence there exists no clique of size ω_f containing only rich vertices. In turn, this implies that $\omega(G) = \omega_f$, since a clique which is not (subset of) a feature clique contains only rich vertices, and we are done because $\omega_f \sim np$ by property \mathcal{A} .

Let us complete the proof by showing that a.a.s. all the events $\mathcal{A}, \mathcal{B}, \mathcal{C}$ hold. Obviously

$$\mathbb{P}[\mathcal{A} \cap \mathcal{B} \cap \mathcal{C}] = 1 - \mathbb{P}[\bar{\mathcal{A}}] - \mathbb{P}[\mathcal{A} \cap \bar{\mathcal{B}}] - \mathbb{P}[\mathcal{A} \cap \mathcal{B} \cap \bar{\mathcal{C}}] \geq 1 - \mathbb{P}[\bar{\mathcal{A}}] - \mathbb{P}[\bar{\mathcal{B}}] - \mathbb{P}[\mathcal{A} \cap \bar{\mathcal{C}}],$$

so it suffices to check that all the probabilities $\mathbb{P}[\bar{\mathcal{A}}], \mathbb{P}[\bar{\mathcal{B}}], \mathbb{P}[\mathcal{A} \cap \bar{\mathcal{C}}]$ tend to zero. For the first two this is immediately implied by Lemma 6.1 (which applies because of $m < n$ and $pn > \ln^4 n$) and Lemma 6.2 respectively. For the latter it follows from Lemma 9.6 and observing that

$$\mathbb{P}[\bar{\mathcal{A}} \cap \mathcal{C}] \leq m \exp\left(-\frac{t^2}{2(1 + \varepsilon)pn \mp t + 2t/3}\right) \stackrel{(9.5)}{\leq} n^{\alpha-1},$$

which does tend to zero, since $\alpha < 1$.

Thus all that remains to be done is to check the two technical observations (9.4) and (9.5). Considering (9.4), we distinguish two cases. For $\sqrt{npm^2} > 3$ we have

$$\begin{aligned} 21 \ln n((1 + \varepsilon)npm^2 + \sqrt{npm^2} \ln n) &\leq 40npm^2 \ln n + 21\sqrt{npm^2} \ln^2 n \\ &= np(40mp \ln n + 21\sqrt{m/n} \ln^2 n). \end{aligned}$$

which is smaller than $(1 - \varepsilon)np$ because of $mp \ll \frac{1}{\ln n}$ and $\alpha < 1$.

And for $\sqrt{npm^2} \leq 3$

$$\begin{aligned} 21 \ln n((1 + \varepsilon)npm^2 + 3 \ln n) &\leq 40npm^2 \ln n + 63 \ln^2 n \\ &\leq 360 \ln^3 n + 63 \ln^2 n. \end{aligned}$$

which is smaller than $(1 - \varepsilon)np$ because of $\frac{\ln^3 n}{n} \ll p$.

Considering (9.5), we distinguish two cases again. For $\sqrt{npm^2} > 3$ we have

$$\begin{aligned} m \exp\left(-\frac{npm^2 \ln^2 n}{2(1 + \varepsilon)npm^2 + \frac{2}{3}\sqrt{npm^2} \ln n}\right) &\leq m \exp\left(-\frac{npm^2 \ln^2 n}{npm^2 \ln n}\right) \\ &= m \exp(-\ln n) = n^{\alpha-1}. \end{aligned}$$

and for $\sqrt{npm^2} \leq 3$

$$\begin{aligned} m \exp \left(-\frac{9 \ln^2 n}{2(1+\epsilon)npm^2 + \frac{2}{3}3 \ln n} \right) &\leq m \exp \left(-\frac{9 \ln^2 n}{100 + 2 \ln n} \right) \\ &\leq m \exp(-\ln n) = n^{\alpha-1}. \end{aligned}$$

□

Chapter 10

Experiments

The main reason to do experiments with our models and algorithms on real-world data is to get a feeling for the appropriateness of the models and the algorithms presented in the chapters before. Are they only of theoretical interest or is it reasonable to apply them? For the models we will see that they are adequate with respect to some parameters while there is much room for improvement, while in the case of the algorithms we have mostly excellent results concerning runtime as well as quality of the results.

We can by no means give a thorough discussion and description of the properties of the networks and can also in most cases give only hints on the reasons why the models and algorithms behave well or not in particular special cases.

10.1 The Giant Component

We tested our result on two instances of complete edge-weighted real world networks on 5119 and 1153 vertices. Here parts of proteins serve as vertices and the edge-weight describes their spatial similarity. If we look at the subgraph of this graph containing all edges with weight greater than a fixed value s (where greater edge weights indicate higher similarity) we can simulate an evolution of this network by gradually decreasing s . Thus first the highly analogue parts get connected and bit by bit also the less similar ones connect to the components.

The evolution found this way differs significantly from a graph in which the same weights are distributed uniformly at random among the edges (see Figure 10.1).

The most striking difference is the slow growth of the largest component in the stages after it has only very few vertices (minimum edge weight between 40 and 60). A similar behaviour cannot be modelled using standard random graphs where \mathcal{N} is either logarithmic or linear in the number of vertices. As one can see in Figure 10.1 the random intersection graph resembles this steady aggregation of vertices to the largest component very well.

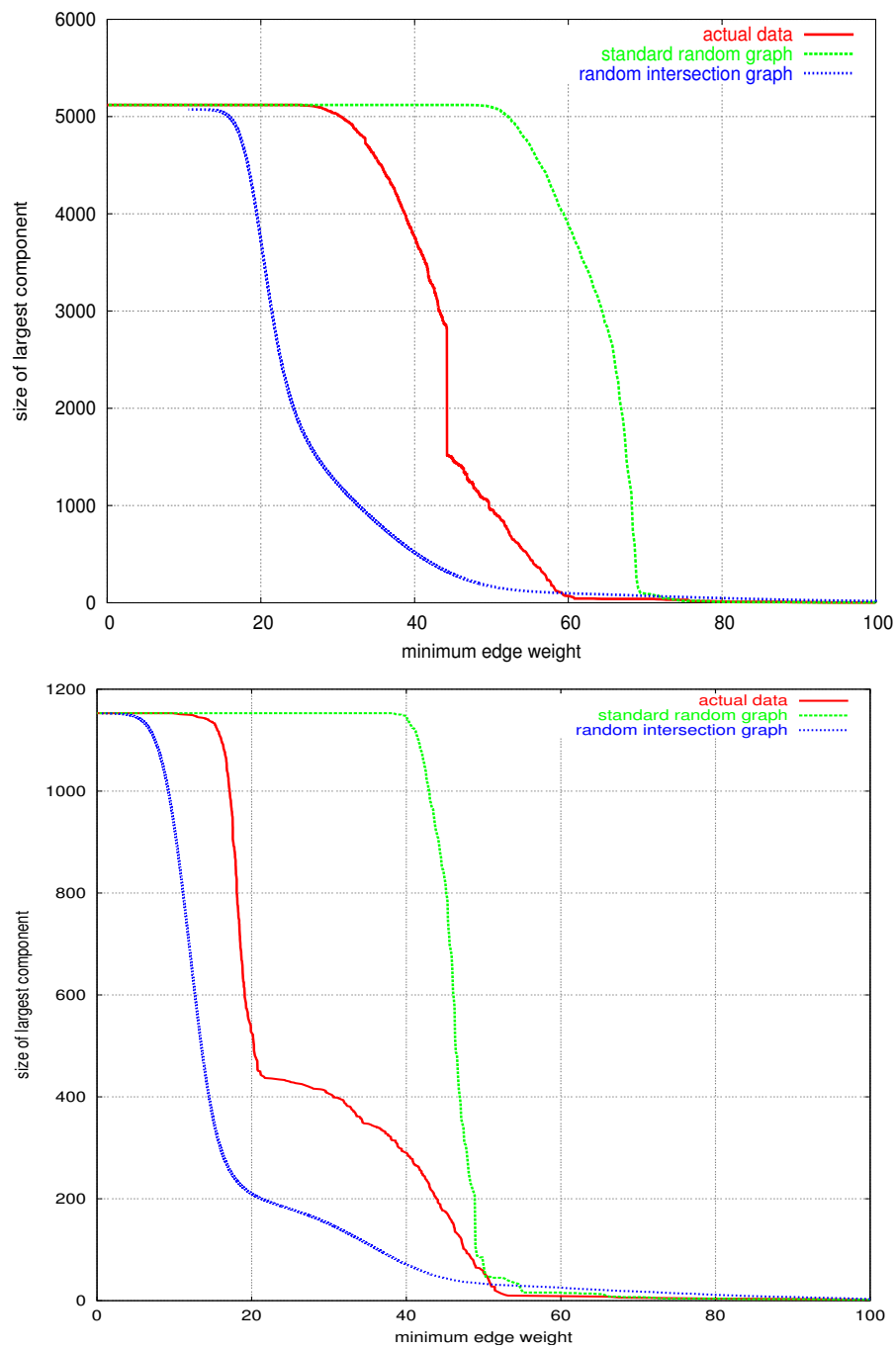


Figure 10.1: Evolution of the largest component in the protein graph.

Type	Name	n	$ I $	$ E $	$ M $
protein structure	1a4j	95	0	213	97
	1AOR	97	0	212	98
	1eaw	53	0	123	56
protein interaction		2113	267	2203	2048
transcription	coli	418	0	519	483
	yeast	688	0	1078	997
molecule similarity	orangebook	2000	1362	163969	
	orangebook2	2000	1254	204614	
	DIP	5119	1429	14434	4054
word cooccurence	bible	9295	31	392066	5195
	darwin	7381	0	44207	16396
	french	8325	0	23841	15977
	japanese	2704	0	7998	4976
	spanish	11586	0	43065	20344
physical internet	Mercator	284805	0	449246	369984
	Internet	75885	0	357317	253578
	cosin	10515	0	21455	14406
	lumeta	209582	0	252714	247960
	opte	40028	0	70979	52055
electronic circuits	s208	122	0	189	171
	s420	252	0	399	363
	s838	512	0	819	747
WWW		325729	0	1090108	431136
social networks	prison	67	0	142	81
	leader2	32	0	80	40
	actor	392340	10121	15038083	94200
	coauthoring	16400	1365	29552	13070

Table 10.1: General statistics on the real-world networks used

10.2 The Networks

We have tested our algorithms on real-world networks from different application areas. The main sources for our networks were Guillaume and Latapy [2004], Albert et al. [2006], Alon [2006]. Table 10.1 gives a brief description of the general characteristics of the network used in terms of number of vertices n , number of isolated vertices $|I|$, number of edges $|E|$ and number of cliques it was generated from (if it was given by cliques) $|M|$.

The protein structure graph represent the adjacency of the secondary-structure elements in some complex proteins according to the PDB database. The protein interaction network gives the undirected view of the network of direct protein interactions in yeast. The transcription networks represent the direct transcription interactions in E. coli and

yeast. The two Orangebook networks are the result of a search for “relatives” of test substances in a database of 2000 drugs where an edge connects a pair of drugs which are relatives to the same test substance. Details concerning this network are described in Thimm et al. [2004]. Moreover “DIP” stands for “Dictionary of Interfaces in Proteins” and is a similarity graph of protein parts (vertices are protein interfaces that are adjacent if they are similar) studied in Frömmel et al. [2003], which is identical to the network studied in the last section when including all edges with weight at least 50.

The word cooccurrence networks describe the adjacency of words (or in the case of the bible their appearance in the same sentence) in texts of various languages. The internet networks describe the structure of the internet either at the level of routers or at the level of autonomous systems collected by different research groups. The electronic circuits networks describe the wired adjacency of different parts of electronic microcircuits and the WWW-graph is a small sample of the world wide web where web pages are vertices and links are edges.

The social networks consist of friendship graphs of prison inmates, students in a course on leadership, costarring graph of actors based on the internet movie database and a coauthoring network of scientific publications.

10.3 Clique Cover

To test the algorithm we started it on each graph with different values of k . It turned out that in the case of very large networks (e.g. the actors graph) even for $k = 2$ it took several days before the algorithm finished, thus we implemented a slightly improved incremental version of Algorithm 1:

Algorithm 4.

Input: Graph $G = (V, E)$ on n vertices

Output: (partial) edge clique cover \mathcal{M} of G

FEATUREFINDINCREMENTAL(G)

```

(1)   $\mathcal{M} := \emptyset;$ 
(2)   $Y := \emptyset;$ 
(3)  foreach  $v \in V$ 
(4)     $Z = N(v)$ 
(5)    if  $G[Z]$  complete
(6)       $\mathcal{M} := \mathcal{M} + Z;$ 
(7)       $Y := Y \cup E(G[Z]);$ 
(8)  repeat
(9)     $Y' := Y$ 
(10)   foreach  $e = (v, w) \in E \setminus Y$ 
(11)      $Z = N(v) \cup N(w)$ 
(12)     foreach  $z \in Z$ 
(13)       if for all  $u \in \Gamma(z) \cap Z$  we have  $(z, u) \in Y$ 
(14)          $Z := Z - z;$ 
(15)       if  $G[Z]$  complete
(16)          $\mathcal{M} := \mathcal{M} + Z;$ 
(17)          $Y := Y \cup E(G[Z]);$ 
(18)  until  $Y' = Y$  or  $Y = E$ 

```

The reasoning behind it is that according to Proposition 8.4 we cannot do wrong if we add a clique to the cover which has an edge in it that is contained in only one inclusion-maximal clique. This statement is still valid if we remove vertices from the common neighbourhood of the edge which are only connected via edges which have already been covered.

The algorithm finishes if we have either no edge which has a clique in the common neighbourhood of its vertices meeting the requirements above or if the whole graph is covered. The number of cliques found this way is still a lower bound to the minimum size of an edge clique cover of the graph, since all cliques are forced somehow (cf. Proposition 8.4).

In two cases we knew in advance the number of features that generated our graph (namely for “Authors” where the publications are the features, and for “Drugs” where the test substances are the features) which should be an upper bound of the number of cliques the algorithm needs to cover the graph.

Table 10.2 gives statistics on the algorithm performance on each graph measured in the number of cliques ($|\mathcal{M}|$) that were needed to cover almost the whole graph (the “coverage” fraction of the edges is given, too) and the values of p and α resulting from this coverage.

As one can see, it is possible to cover a large portion of the graph with a number of cliques that is relatively smaller than the number of edges and also smaller than the number of cliques needed by the algorithm in Guillaume and Latapy [2004] (which covered the whole graph).

In order to give further evidence for the adequacy of our model we compared the de-

Name	n	$ E $	$ M $	α	$\log_n p$	coverage
1a4j	95	213	97	1	-0.83	100
1AOR	97	212	98	1	-0.83	100
leaw	53	123	56	1.01	-0.81	100
proteins	2113	2203	2048	0.99	-0.94	100
coli	418	519	483	1.02	-0.93	100
yeast	688	1078	997	1.05	-0.94	100
orangebook	2000	163969	12	0.32	-0.32	100
orangebook2	2000	204614	20	0.39	-0.34	100
DIP	5119	14434	4054	0.97	-0.88	86.7
bible	9295	392066	5195	0.93	-0.72	70.4
darwin	7381	44207	16396	1.08	-0.9	83.7
french	8325	23841	15977	1.07	-0.93	97.7
japanese	2704	7998	4976	1.07	-0.92	96.9
spanish	11586	43065	20344	1.06	-0.92	91.4
Mercator	284805	449246	369984	1.02	-0.96	97.9
Internet	75885	357317	253578	1.1	-0.95	97
cosin	10515	21455	14406	1.03	-0.94	99.4
lumeta	209582	252714	247960	1.01	-0.97	99.8
opte	40028	70979	52055	1.02	-0.95	99.9
s208	122	189	171	1.07	-0.91	100
s420	252	399	363	1.06	-0.92	100
s838	512	819	747	1.06	-0.93	100
www	325729	1090108	431136	1.02	-0.93	91.8
prison	67	142	81	1.04	-0.84	100
leader2	32	80	40	1.06	-0.79	100
actor	392340	15038083	94200	0.88	-0.77	96.5
coauthoring	16400	29552	13070	0.97	-0.92	99.9

Table 10.2: Statistics on the performance of Algorithm 4 on real-world networks

gree distribution for small degrees of some original real-world networks and our theoretical prediction based on the degree distribution of random intersection graphs calculated in Chapter 7. The results are shown in Figure 10.2.

Especially for smaller graphs and smaller degrees the approximation is quite good. Of course it is not quite as good as that in Guillaume and Latapy [2004], but this is due to the fact that there the whole degree distribution was used as an input, whereas we only have the two parameters p and m to adjust the model.

For the Orangebook test set the theoretical predicted degrees are smaller than the experimental ones. This is due to the so-called “bipartite clustering” (as described in Guillaume and Latapy [2004]) which in our case means that the features are not completely independent but somewhat transitive, as there are “similar” features. This results in a larger overlap between some feature cliques than is theoretically predicted and thus leads to larger degrees of the vertices involved.

Furthermore there is clearly the effect of the exponential cutoff in all of the degree distributions, which results from the effect that the real-world networks are expected to have a power-law degree distribution while it was shown that random intersection graphs exhibit an exponential cutoff Stark [2004].

10.4 Colouring

In Table 10.3 Greedy χ , GreedyPES χ and GreedySL χ denote the number of colours needed by a greedy colouring procedure that colours the vertices in the natural order (in which they were read), in a PES ordering (cf Algorithm 2) and in a smallest last ordering (cf Algorithm 3) respectively. Table 10.1 also states the size of the largest clique we were able to find in the graphs using the clique cover algorithm described in Chapter 8 and in the cases where we did not colour optimally also the largest clique found by enumeration methods (in brackets). Obviously the difference between the proposed number of colours and the proposed size of a largest clique is an upper bound of the distance of either number to the optimal value.

The results show that the colouring algorithms seem to perform well on real-world graphs. In all cases of biological networks (and in two thirds of all the cases) we were able to colour the graph optimally using the heuristic described in Algorithm 3.

We also performed an additional test to obtain some indication as to how difficult it really is to optimally colour these particular input graphs. For this, we determined the so-called k -core by repeatedly removing all vertices with degree smaller than k , where we set k as the size of the largest known clique. If the k -core were very small or of a simple structure for which one could easily find a k -colouring, then it would be trivial to extend this colouring to a valid and thus optimal k -colouring of the whole graph by re-attaching the vertices in reverse order. (Note that this procedure is essentially identical to Algorithm 3, except that it is forced to stop when it realizes that all remaining vertices $x \in A$ satisfy $|\Gamma(x) \cap A| \geq k$.) However, as shown in Table 10.1, in many cases the size of the k -core is substantially larger than that of the largest known clique.

Finally, we remark that the large difference between the proposed colouring number

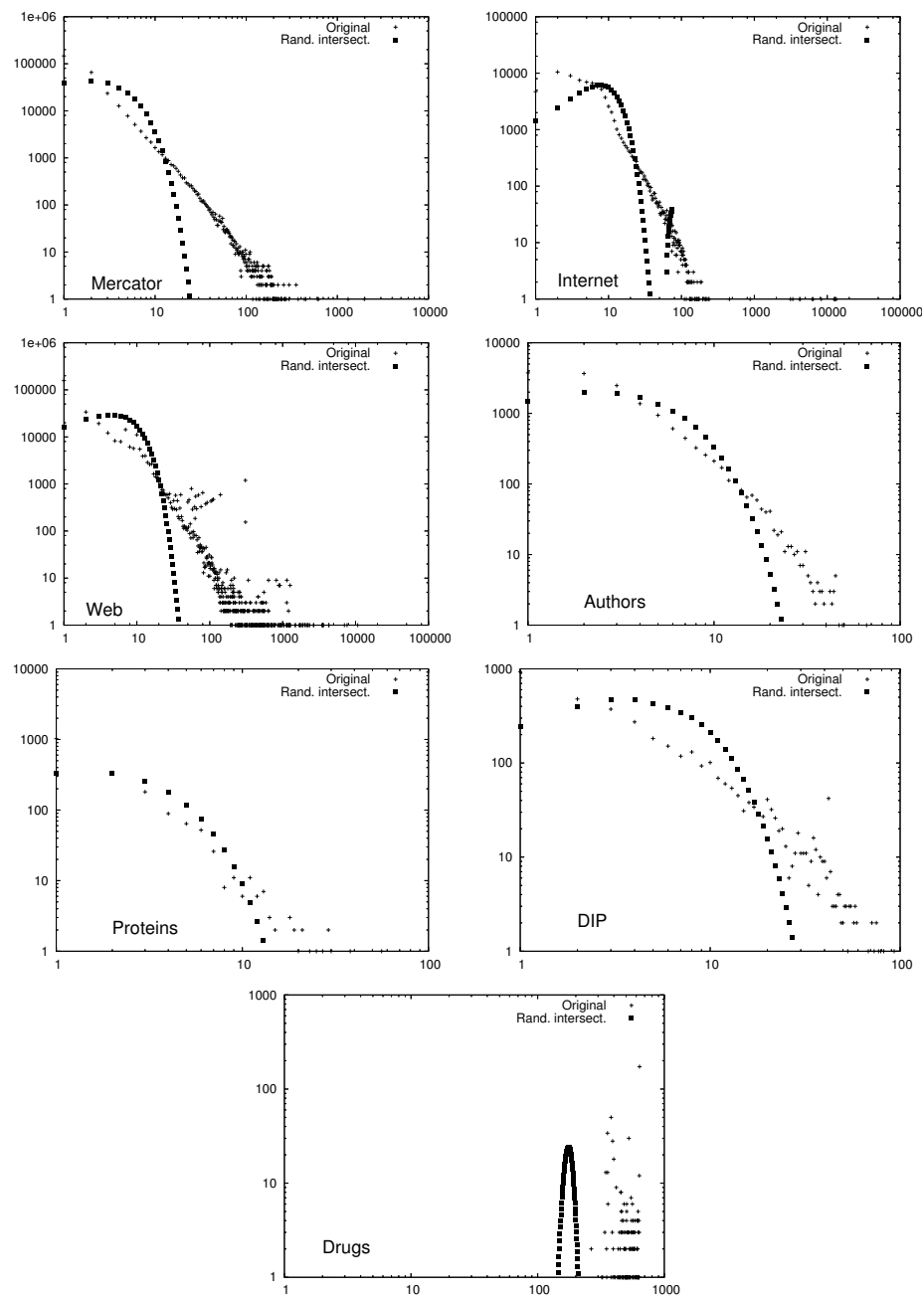


Figure 10.2: Degree distributions for real-world networks: experimental results and theoretical predictions

Name	χ Greedy	χ PES	χ SL	clique	core
1a4j	5	5	4	4	0
1AOR	6	5	5	5	0
leaw	5	5	5	5	0
proteins	6	6	6	6	0
coli	4	4	3	3	52
yeast	5	4	4	4	0
orangebook	381	381	381	381	432
orangebook2	384	381	381	381	555
DIP	42	42	42	42	0
bible	143	117	118	39 (90)	3176
darwin	23	20	20	11 (16)	1392
french	14	11	12	8 (8)	709
japanese	12	11	10	7 (9)	382
spanish	21	18	17	10 (14)	1194
Mercator	38	36	33	13 (27)	1453
Internet	22	21	20	18 (20)	996
cosin	16	17	16	14 (16)	74
lumeta	10	8	8	8	150
opte	8	7	8	6 (6)	98
s208	4	4	3	3	0
s420	4	4	3	3	0
s838	4	4	3	3	0
www	155	155	155	155	1367
prison	6	5	5	5	0
leader2	5	4	4	4	17
actor	294	294	294	294	2647
coauthoring	11	8	8	8	0

Table 10.3: Statistics on the performance of the colouring algorithms on real-world networks

and the proposed clique number for the internet and word cooccurrence networks is not so much a failure of the colouring algorithms. Instead, it seems mainly due to the fact that the clique cover algorithm, described in Chapter 8 with the aim to find a good clique *cover*, cannot find a large clique on those instances – a simple enumeration method applied to higher cores of the graphs often identified larger cliques (see the numbers in brackets).

Chapter 11

Conclusion and Outlook

11.1 Random Intersection Graphs

We have seen that random intersection graphs while not covering all aspects of real-world networks give a good starting point for a semantic analysis of the structure of those networks. The straightforward model of a uniform probability for each feature, to be chosen by a vertex gives a means of analyzing graph evolutions which have building blocks (in this case the feature cliques) which grow during the evolution of the graph. This is also the essential difference to the study of random hypergraphs.

Random hypergraphs often give – due to their limitation to a constant edge size – a result which is much closer to the classical Erdős-Rényi-graphs than to a real-world application. However applying appropriate (probability) distributions to the size of the feature sets (or the feature cliques) as proposed by Godehardt and Jaworski [2001] might give better approximations of real-world data although first studies by Jaworski et al. [2006] and Rybarczyk [2006] point more in the direction of the equivalence to classical random hypergraphs.

Another extension of the model comes from the idea of a varying overlap in the feature sets, i.e. it is necessary to have at least l features in common in order to create an edge between two vertices. While a constant l may create no conceptual difference in the case of the random intersection graphs studied here, they will certainly make a difference if the size of the feature sets is limited by a constant as well. Furthermore growing l will lead to new effects in the study of the binomial model.

Another idea with regards to extensions of the model is the introduction of meta-features. That is we first put the features into groups according to some probability distribution and let the vertices first choose one (or several) of those groups (meta-features) and then select individually with a different probability the features from the group. This would account for the fact that in most cases the features are not independent e.g. if the proposed features are known in advance it is often the case that the feature cliques tend to overlap more than in the theoretical modelling. This effect of positive correlation between features is also called bipartite clustering.

Furthermore one could introduce an edge probability which is proportional to the

number of features two vertices have in common to account for the fact that in the real-world networks not all edges may be present due to errors in the data.

Two further parameters of real-world networks which attracted considerable interest are the diameter (or rather the average distance between vertices) of the graph and the so-called clustering coefficient. For some ideas on how to tackle them see Sections 11.4 and 11.5.

11.2 Clique Cover

Our analysis yields a rigorous proof for the asymptotic optimality of our simple greedy algorithm in the random intersection graph model $G_{n,m,p}$ for a certain range of m and p . Experimental results indicate that even outside this range the algorithm performs well, for example when $\alpha > 1$. It is clear that the reconstruction of feature cliques becomes impossible once they are no longer maximal, which seems to happen when p is of order $m^{-1/2}$. It would be interesting to prove that this (or a different) algorithm succeeds up to this point.

Furthermore one could easily extend the algorithm not to find *clique* covers but to cover the graph with dense subgraphs only. This would on one hand correspond to a different intersection graph model, where overlapping feature sets of two vertices do not imply directly an edge between the vertices but only increase the edge probability and on the other hand to the experimental fact of noisy or erroneous data omitting edges which “should be there”.

11.3 Colouring and Independence Number

For the ranges not covered by Theorems 9.1 and 9.2, the chromatic number seems to be more difficult to estimate. From the aforementioned result by Singer [1995] it is clear that those graphs are no longer chordal for $p \gg \sqrt{\frac{1}{nm}}$ while the results on the clique cover in Chapter 8 suggest that the feature cliques stay the dominant structural element up to $p < \min\{\frac{1}{5}m^{-\frac{2}{3}}, \frac{n}{8m^2}\}$.

In higher ranges, the approximation of the chromatic number by the size of the largest feature clique will not be very good. Using a different approach, Ueckerdt [2005] tried to establish a better lower bound via the independence number. Using the fact that the chromatic number of any graph is at least as high as the number of vertices divided by the size of a largest independent set, we obtain a lower bound on the chromatic number which beats the size of the largest feature clique, as the following result shows.

Theorem 11.1 (Ueckerdt [2005]). *Let $\varepsilon > 0$ be fixed and let $m := n^\alpha$ with $\alpha > 0$ fixed and $\frac{\ln n}{m} \ll p \ll \sqrt{\frac{\ln n}{m}}$. Then a.a.s. the random intersection graph $G_{n,m,p}$ has no independent set of size*

$$(2 + \varepsilon) \frac{\ln n}{mp^2},$$

which implies that

$$\chi(G_{n,m,p}) \geq \frac{p^2 mn}{(2 + \varepsilon) \ln n} \gg pn.$$

Lower bounds on the independence number (which match the upper bounds by a logarithmic factor) can also be found in Ueckerdt [2005].

11.4 Diameter

The diameter of real-world networks was found to be quite small which is not surprising to people familiar with the theory of random graphs where there a lot of results on graphs having diameter logarithmic in the number of vertices (see Bollobás and Riordan [2002]).

The diameter of random bipartite graphs was already studied by Bollobás and Klee [1984] and their result obviously gives an upper bound to the diameter of random intersection graphs, since the the diameter of an intersection graph is at most half of the diameter of its bipartite generator.

This gives rise to the following theorem which is a corollary to Theorem B in Bollobás and Klee [1984].

Theorem 11.2. *Let m, n and p be such that $pn \geq pm \geq \ln^4 n$ and let k be a fixed positive integer. If*

$$p^{2k+1} m^k n^k \gg \ln(mn)$$

then a.a.s. $\text{diam}(G_{n,m,p}) \leq k + 1$.

11.5 Clustering Coefficient

The term “clustering coefficient” is used ambiguously in the literature. In general it should describe the tendency of a graph to cluster, i.e. to have rather cliques than trees as induced subgraphs. One possiblity to measure this which is used on a per vertex basis is to calculate for a vertex the number of edges appearing in its neighborhood in relation to the number of possible edges. The clustering coefficient of a graph is then the average cluster coefficient of its vertices.

$$\text{cc}_1(v) := \frac{\left| \binom{\Gamma(v)}{2} \cap E \right|}{\left| \binom{\Gamma(v)}{2} \right|} \quad \text{cc}_1(G) := \frac{\sum_{v \in V} \text{cc}_1(v)}{|V|}$$

A second possibility to define this coefficient is to divide three times the number of all triangles in the graph by the number of all paths of length 2 (i.e. three vertices connected by two edges).

$$\text{cc}_2(G) := \frac{3 \# K_3 \subseteq G}{\# P_2 \subseteq G} \tag{11.1}$$

which can also be stated as a weighted sum in terms of $cc_1(v)$

$$cc_2(G) = \frac{\sum_{v \in V} cc_1(v) \left| \binom{\Gamma(v)}{2} \right|}{\sum_{v \in V} \left| \binom{\Gamma(v)}{2} \right|}$$

It was already pointed out by Bollobás and Riordan [2002] that the two parameters can differ by a factor linear in the number of vertices.

One straightforward result on the clustering coefficient of random intersection graphs is the following theorem

Theorem 11.3. *Let m, n with $m = n^\alpha$ and p be such that*

$$p \ll \begin{cases} n^{-\frac{1}{2}} m^{-1} & \text{for } \alpha \leq \frac{1}{2} \\ n^{-\frac{3}{4}} m^{-\frac{1}{2}} & \text{for } \alpha \geq \frac{1}{2} \end{cases}$$

Then a.a.s. $cc_1(G_{n,m,p}) = cc_2(G_{n,m,p}) = 1$.

(Maybe it is worthwhile to note that the graph may in fact contain edges and even triangles at this stage of evolution.) The reason for the large clustering coefficient is simply that the values above are the thresholds for the appearance of an induced P_2 in a random intersection graph as shown in Singer [1995].

The definition (11.1) is more appealing to the random graph theorist since it gives precisely the probability of an edge to close a triangle conditioned on the fact that the two other edges are already there.

Theorem 11.4. *Let m, n with $m = n^\alpha$ and p be such that $p^2 m \rightarrow 0$. Then*

$$cc_2(G_{n,m,p}) \sim \frac{mp^3 + (mp^2)^3}{mp^3 + (mp^2)^2}$$

For a proof see the results on the subgraphs of $G_{n,m,p}$ in Singer [1995].

This theorem gives a good idea of where the large clustering coefficient comes from. mp^3 is approximately the probability that the three vertices are joined because they see the same feature while mp^2 is the probability for a single edge induced by a feature. Thus as long as the first summand dominates, the clustering coefficient is close to 1 while if the second summand dominates the clustering coefficient approaches $p^2 m$ (which is equivalent to the edge probability and thus to $G_{n,p'}$ with $p' = p^2 m$).

11.6 Final Remarks

This thesis gives a brief account of the search for models for real-world networks arising in the life sciences and other areas. Our tools come mainly from the theory of random graphs which seem well suited at first sight (what if not a graph should model a network) but also show the necessity to include further aspects of the real-world networks into the models. Despite of the striking similarities of the networks studied (concerning for

instance clustering) it became also clear that applicationwise there are differences (see for instance the results on colouring) and we cannot hope for the grand unified model. Nevertheless there are still a lot of (optimization) problems to be solved on large real-world instances (see for instance Henzinger [2003] which require further analysis of the models as well as development and analysis of new algorithms).

Bibliography

- R. Albert and A.-L. Barabási. Statistical mechanics of complex networks. *Reviews of Modern Physics*, 74(47), 2002. URL <http://arxiv.org/abs/math/0106096>.
- R. Albert, H. Jeong, and A.-L. Barabási. Database of self-organized networks, 2006. <http://www.nd.edu/networks/database/index.html>.
- U. Alon. Collection of complex networks from different application areas, 2006. <http://www.weizmann.ac.il/mcb/UriAlon/groupNetworksData.html>.
- T. Andriamampianina and V. Ravelomanana. Enumeration of connected uniform hypergraphs. In *Proceedings of FPSAC 2005*, 2005.
- K. B. Athreya and A. N. Vidyashankar. Branching processes. Technical report, University of Georgia – Department of Statistics, 1999.
- A. D. Barbour, M. Karoński, and A. Ruciński. A central limit theorem for decomposable random variables with applications to random graphs. *Journal of Combinatorial Theory (B)*, 47:21, 1989.
- D. Barraez, S. Boucheron, and W. de la Vega. On the fluctuations of the giant component. *Combinatorics, Probability and Computing*, 9:287–304, 2000.
- M. Behrisch. Component evolution in random intersection graphs. *Electronic Journal of Combinatorics*, 2006. to appear.
- M. Behrisch and A. Taraz. Efficiently covering complex networks with cliques of similar vertices. *Theoretical Computer Science*, 355(1):37–47, 2006.
- M. Behrisch, A. Taraz, and M. Ueckerdt. Colouring random intersection graphs and complex networks. Preprint, December 2005.
- E. Bender, E. Canfield, and B. McKay. The asymptotic number of labeled connected graphs with a given number of vertices and edges. *Random Structures and Algorithms*, 1:127–169, 1990.
- E. Bender, E. Canfield, and B. McKay. Asymptotic properties of labeled connected graphs. *Random Structures and Algorithms*, 3:183–202, 1992.

- B. Bollobás. *Random Graphs*. Cambridge University Press, 2nd edition, 2001.
- B. Bollobás and V. Klee. Diameters of random bipartite graphs. *Combinatorica*, 4(1): 7–19, 1984.
- B. Bollobás and O. Riordan. Mathematical results on scale-free random graphs. In H. Schuster and S. Bornholdt, editors, *Handbook of graphs and networks*, chapter 1. Wiley-VCH, first edition, 2002.
- B. Bollobás and O. Riordan. Slow emergence of the giant component in the growing m-out graph. to appear in *Random Structures and Algorithms*, 2005.
- F. Chung and L. Lu. Connected components in random graphs with given expected degree sequences. *Annals of Combinatorics*, 6:125–145, 2002.
- A. Coja-Oghlan, C. Moore, and V. Sanwalani. Counting connected graphs and hypergraphs via the probabilistic method. *Random Structures and Algorithms*, 2006. URL <http://www.informatik.hu-berlin.de/~coja/jhyper5.ps>. to appear.
- D. Coppersmith, D. Gamarnik, M. Hajiaghayi, and G. Sorkin. Random MAX SAT, random MAX CUT, and their phase transitions. *Random Structures and Algorithms*, 24:502–545, 2004.
- R. Diestel. *Graph theory*. Springer, New York, 1997.
- P. Erdős and A. Rényi. On random graphs i. *Publicationes Mathematicae Debrecen*, 5: 290–297, 1959.
- P. Erdős and A. Rényi. On the evolution of random graphs. *Publ. Math. Inst. Hung. Acad. Sci.*, 5:17–61, 1960.
- J. A. Fill, E. R. Scheinerman, and K. B. Singer-Cohen. Random intersection graphs when $m = \omega(n)$: An equivalence theorem relating the evolution of the $G(n, m, p)$ and $G(n, p)$ models. *Random Structures and Algorithms*, 16(2):156–176, March 2000.
- C. Frömmel, C. Gille, A. Goede, C. Gröpl, S. Hougardy, T. Nierhoff, R. Preissner, and M. Thimm. Accelerating screening of 3D protein data with a graph theoretical approach. *Bioinformatics*, 19(18):2442–2447, 2003.
- M. R. Garey and D. S. Johnson. *Computers and Intractability*. W.H. Freeman and Company, 1979.
- E. Godehardt and J. Jaworski. Two models of random intersection graphs and their applications. *Electronic Notes in Discrete Mathematics*, 10, 2001. URL <http://www1.elsevier.com/gej-ng/31/29/24/49/27/61/endm10036.ps>.
- J.-L. Guillaume and M. Latapy. Bipartite structure of all complex networks. *Information Processing Letters*, 90:215–221, 2004.

- M. R. Henzinger. Algorithmic challenges in web search engines. *Internet Mathematics*, 1(1):115–126, 2003.
- S. Janson, T. Łuczak, and A. Ruciński. *Random Graphs*. John Wiley & Sons, 2000.
- J. Jaworski, M. Karonski, and D. Stark. The degree of a typical vertex in generalized random intersection graph models. *Discrete Mathematics*, 306:2152—2165, 2006.
- M. Karoński and T. Łuczak. The phase transition in a random hypergraph. *J. Comput. Appl. Math.*, 142:125–135, 2002.
- M. Karoński and T. Łuczak. The number of connected sparsely edged uniform hypergraphs. *Discrete Math.*, 171:153–168, 1997.
- M. Karoński, E. R. Scheinerman, and K. B. Singer-Cohen. On random intersection graphs: The subgraph problem. *Combinatorics, Probability and Computing*, 8:131–159, 1999.
- T. Łuczak. On the number of sparse connected graphs. *Random Structures and Algorithms*, 1:171–173, 1990.
- M. E. J. Newman, S. H. Strogatz, and D. J. Watts. Random graphs with arbitrary degree distributions and their applications. *Physical Review E*, 64, 2001. URL <http://aps.arxiv.org/abs/cond-mat/0007235/>.
- N. O’Connell. Some large deviation results for sparse random graphs. *Prob. Th. Relat. Fields*, 110:277–285, 1998.
- B. Pittel. On tree census and the giant component in sparse random graphs. *Random Structures and Algorithms*, 1(3):311–342, 1990.
- B. Pittel and N. Wormald. Asymptotic enumeration of sparse graphs with a minimum degree constraint. *J. Combin. Theory, Series A*, 101:249–263, 2003.
- B. Pittel and N. Wormald. Counting connected graphs inside out. *J. Combin. Theory, Series B*, 93:127–172, 2005.
- V. Ravelomanana and A. Rijamamy. Creation and growth of components in a random hypergraph process. Preprint, 2005.
- K. Rybarczyk, 2006. Personal Communication.
- K. Rybarczyk. On some applications of random hypergraphs (polish). Master’s thesis, Adam Mickiewicz University, Poznan, 2005.
- E. R. Scheinerman. Random interval graphs. *Combinatorica*, 8(4):357–371, 1988.
- J. Schmidt-Pruzan and E. Shamir. Component structure in the evolution of random hypergraphs. *Combinatorica*, 5:81–94, 1985.

- K. B. Singer. *Random Intersection Graphs*. PhD thesis, John Hopkins University, Baltimore, Maryland, 1995.
- D. Stark. The vertex degree distribution of random intersection graphs. *Random Structures and Algorithms*, 24(3):249–258, May 2004.
- C. Stein. A bound for the error in the normal approximation to the distribution of a sum of dependent variables. In *Proc. 6th Berkeley Symposium on Mathematical Statistics and Probability*, pages 583–602, 1970.
- V. E. Stepanov. On the probability of connectedness of a random graph $G_m(t)$. *Theory Prob. Appl.*, 15:55–67, 1970.
- R. E. Tarjan and M. Yannakakis. Simple linear-time algorithms to test chordality of graphs, test acyclicity of hypergraphs, and selectively reduce acyclic hypergraphs. *SIAM Journal on Computing*, 13(3):566–579, 1984.
- M. Thimm, A. Goede, S. Hougardy, and R. Preissner. Comparison of 2D similarity and 3D superposition. application to searching a conformational drug database. *Journal of Chemical Information and Computer Sciences*, 44:1816–1822, 2004.
- M. Ueckerdt. Färben von zufälligen Schnittgraphen. Diploma thesis, Humboldt-Universität zu Berlin, 2005.
- R. van der Hofstad and J. Spencer. Counting connected graphs asymptotically. Preprint. To appear in *European Journal on Combinatorics.*, 2005.

Notation

Symbol	Description
$G = (V, E)$	graph G with vertex set V and edge set E
$\Gamma(v)$	neighborhood of a vertex v (set of adjacent vertices)
$N(v)$	inclusive neighborhood of a vertex v ($N(v) := \Gamma(v) + v$)
$d(v)$	degree of a vertex v ($d(v) := \Gamma(v) $)
$G_{n,p}$	binomial random graph (Erdős-Rényi-model)
$G_{n,m}$	uniform random graph (Erdős-Rényi-model)
$H_d(n, p)$	binomial random d -uniform hypergraph
$H_d(n, m)$	uniform random d -uniform hypergraph
$G_{n,m,p}$	random intersection graph
$B_{n,m,p}$	random bipartite graph
$\mathcal{N}(G)$	order of the largest component of G (number of vertices)
$\mathcal{M}(G)$	size of the largest component of G (number of edges)
$\omega(G)$	size of the largest clique in G
$\omega_f(G)$	size of the largest feature clique in G (intersection graph)
$\chi(G)$	chromatic number of G
$\chi_A(G)$	number of colours used by algorithm A to colour G
$\text{diam}(G)$	diameter of G
$\text{cc}(G)$	clustering coefficient of G

Acknowledgement

First of all I would like to thank Professor Hans Jürgen Prömel, for giving me the chance to pursue the research project leading to this thesis inside the inspiring and supporting MATHEON research center. I am also indebted to him for guiding my interest into the direction of random graphs by introducing me to Anusch Taraz without whom I would not have spent a second day thinking about these problems. To him go my deepest thanks also for the nice way of collaboration we both could follow leading to most of the results in Part II of this thesis.

Furthermore I would like to thank Amin Coja-Oghlan for having the patience to wait and explain over and over again until I finally grasped the important parts of the limit theorems and their applications, as well as Mihuyun Kang, for a lot of proofreading and fruitful discussions. Without them the world would still wait for the results in Part I to come into existence.

Finally I would like to thank the authors of Frömmel et al. [2003], Thimm et al. [2004], Guillaume and Latapy [2004], Alon [2006] for generous access to their databases, which made it possible to perform the experiments on a somehow meaningful set of information.

Lebenslauf

Persönliche Daten

Name Michael Behrisch

Geburtsdatum 13. Juli 1976

Geburtsort Berlin

Adresse Eschenbachstr. 8
12437 Berlin
Tel.: (030) 345 03 555

E-Mail michael.behrisch@web.de

Ausbildung

06/1996 Abitur an der Heinrich-Hertz-Oberschule (Gymnasium) in Berlin

10/1997–06/2002 Studium der Informatik an der Humboldt-Universität zu Berlin

06/2002 Diplom Informatik, Nebenfächer Physik und Philosophie

Berufstätigkeit

07/1998–05/2001 Werkstudent in der Forschung der DaimlerChrysler AG in Berlin

06/2001–07/2002 Webadministrator der Forschergruppe „Algorithmen, Struktur, Zufall“ am Institut für Informatik der Humboldt-Universität zu Berlin

07/2002–09/2006 wissenschaftlicher Mitarbeiter im Teilprojekt A5 „Analyse und Modellierung komplexer Netzwerke“ des DFG-Forschungszentrums MATHEON an der Humboldt-Universität zu Berlin

Berlin, den 19. Januar 2008

Michael Behrisch

Erklärung

Ich erkläre hiermit, dass

- ich die vorliegende Dissertationsschrift selbstständig ohne fremde Hilfe verfasst und nur die angegebene Literatur und Hilfsmittel verwendet habe,
- ich mich nicht bereits anderwärts um einen Doktorgrad beworben habe oder einen solchen besitze und
- mir die Promotionsordnung der Mathematisch-Naturwissenschaftlichen Fakultät II der Humboldt-Universität zu Berlin bekannt ist.

Berlin, den 19. Januar 2008

Michael Behrisch